

Uncovering and Popularizing the Genomic Mosaic of Grain Amaranth Domestication



Inaugural-Dissertation

zur Erlangung des Doktorgrades der Mathematisch-Naturwissenschaftlichen Fakultät
der Universität zu Köln vorgelegt von

José Miguel Gonçalves Dias

aus Fafe, Portugal

Köln, 2024



Berichterstatter: Prof. Dr. Markus G. Stetter
(**Gutachter**) Prof. Dr. Michael Nothnagel

Tag der mündlichen Prüfung: 12.06.2024

Zusammenfassung

Die Evolution von Kulturpflanzen, insbesondere von Amaranth, wurde lange Zeit als lineare Entwicklung von der Wildform zur domestizierten Pflanze angesehen. Die unvollständige Domestizierung von Amaranth und seine lange Anbaugeschichte machen ihn jedoch zu einem hervorragenden Modell für die Untersuchung der komplexen Genomdynamik von Nutzpflanzen. Ziel dieser Arbeit ist es, das komplizierte genomische Mosaik des Domestikationsprozesses bei fünf wilden und domestizierten Amaranth-Unterarten zu entschlüsseln.

Mithilfe von Populationsgenomik und Computerbiologie werden in dieser Studie die genetischen Auswirkungen auf Amaranth-Populationen untersucht, einschließlich des Genflusses und des Potenzials für die Artbildung. Das Hauptziel dieser Arbeit ist es, die genomischen Auswirkungen des Kontakts nach der Domestizierung auf fünf Amaranth-Populationen in Amerika zu untersuchen, die sowohl wild als auch domestiziert sind. Die Studie zeigt, dass Amaranth-Pflanzenarten nach der Domestizierung nicht nur genetisches Material austauschen, sondern dass der Kontakt mit ihrem wilden Vorfahren auch dazu beiträgt, ihre genetische Vielfalt und damit ihre evolutionäre Rettung zu erhalten.

Darüber hinaus werden im Rahmen dieser Forschung computergestützte Werkzeuge und Ressourcen für die Verarbeitung und Analyse der riesigen Menge an generierten Genomdaten entwickelt. In dieser Arbeit wird PopAmaranth vorgestellt, ein neuartiger Genom-Browser für die Populationsgenetik. "PopAmaranth bietet eine intuitive Schnittstelle, die verschiedene Funktionen wie Selektionssignale, Genannotation und Variantenauftrufe integriert und es interdisziplinären Forschern ermöglicht, genomische Daten im Populationsmaßstab zu erforschen und Untersuchungen in der Populationsgenomik und Pflanzenzüchtung zu erleichtern."

Zusammenfassend gewährt diese Doktorarbeit neue Einsichten in die evolutionäre Geschichte von Amaranth und verdeutlicht das komplizierte Verhältnis von domestizierten und wilden Populationen. Die Forschungsarbeit unterstreicht die Bedeutung von Comput-

erprogramme wie Genom-Browsern für die Erleichterung von Genomanalysen auf Populationsebene. Diese Ergebnisse tragen zum Bereich der Populationsgenomik bei und fördern unser Verständnis der genetischen Dynamik, die der Diversifizierung von Arten zugrunde liegt.

Abstract

Crop evolution, especially that of amaranth, has long been thought to be a linear progression from wild to domesticated. However, amaranth's incomplete domestication and extensive cultivation history make it an excellent model for studying the complex genome dynamics of crops. This thesis aims to unravel the intricate genomic mosaic of the domestication process in five amaranth sub-species, wild and domesticated.

Through population genomics and computational biology, this study investigates the genetic impact on amaranth populations, including gene flow and the potential for speciation. The primary objective of this thesis is to explore the genomic effects of post-domestication contact on five amaranth populations, both wild and domesticated, found in the Americas. The study shows that not only do amaranth crop species exchange genetic material after domestication, but contact with their wild ancestor also contributes to maintaining their genetic diversity and, consequently, evolutionary rescue.

Moreover, this research creates computational tools and resources to handle and analyze the vast amount of genomic data generated. The thesis introduces PopAmaranth, a novel genome browser for population genetics. PopAmaranth provides an intuitive interface integrating various features such as selection signals, gene annotation, and variant calls, enabling interdisciplinary researchers to explore population-scale genomic data and facilitate investigations in population genomics and crop breeding.

In summary, this doctoral thesis provides novel insights into the evolutionary history of amaranth, unraveling the intricate interplay between domesticated and wild populations. The research highlights the importance of computational tools like Genome Browsers in facilitating population-scale genomic analyses. These findings contribute to the field of population genomics, advancing our understanding of the genetic dynamics underlying species diversification.

Contents

1	Introduction	3
1.1	Genomics and Computational Biology enable Population Genomics	3
1.1.1	Resources in Computational Biology increase accessibility of Ge- nomics results	4
1.2	Population Genomics to study the forces driving evolution	5
1.2.1	Domestication as framework for evolution	8
1.3	Revolution in non-model plant research	9
1.4	Amaranth as a model system for studying genomic mosaic of domestication	10
1.5	Aims of the Thesis	12
2	Genetic incompatibilities and evolutionary rescue by wild relatives shaped grain amaranth domestication	14
2.1	Introduction	14
2.2	Results	18
2.2.1	Fine-scale gene flow reveals diverse local ancestry of grain amaranths	19
2.2.2	Introgression from wild ancestor mitigates increased genetic load in domesticated grain amaranths	21
2.2.3	Hybrid incompatibilities between grain amaranths	22
2.3	Discussion	25
2.4	Materials and Methods	27
2.4.1	Demographic modeling	29
2.5	Data Availability	31
2.6	Author contribution	31
2.7	Competing interests	32
2.8	Acknowledgments	32

3 PopAmaranth: a population genetic genome browser for grain amarants and their wild relatives	33
3.1 Introduction	33
3.2 Methods	35
3.2.1 Data and filtering	35
3.2.2 Population genetic browser tracks	35
3.2.3 Browser implementation and annotation	36
3.2.4 PopAmaranth application to candidate genes	36
3.3 Results	37
3.3.1 Sample filtering	37
3.3.2 Categories and Tracks	37
3.3.3 PopAmaranth case study	40
3.4 Discussion	41
3.5 Availability	42
3.6 Acknowledgments	42
4 Overarching Discussion	43
5 Concluding Remarks	48
6 Bibliography	49
A Supplementary Information Chapter 2	69
B Supplementary Information Chapter 3	83
Acknowledgments	97
Erklärung zur Dissertation	98
Curriculum Vitae	99

List of Figures

1	Conceptual Overview	12
2	Genome-wide signals of gene flow.	17
3	Variable ancestry across individuals and along the genome.	19
4	Fine-scale gene flow along the genome between domesticated amaranth populations.	20
5	Genetic load in domesticated and wild amaranth.	23
6	Proportion of lethal F_1 phenotypes.	24
7	Principal Component Analysis with filtered samples.	37
8	PopAmaranth screen view	40
9	Conceptual Framework Outcome	45
S1	Occurrence map for two wild and three domesticated species of grain ama- ranth	69
S2	Different demographic models used in Fastsimcoal2 to predict the best scenario.	70
S3	D-statistic value for comparisons between individuals.	71
S4	D-statistic value for comparisons between populations of South America.	71
S5	Ancestry proportions summary along the genome, per recipient population, per scaffold.	72
S6	Summary for multiple genome alignment and GERP score.	73
S7	Distribution of genetic load during domestication of grain amaranth using two different genomes for accounting for reference bias.	74
S1	PCA before filtering.	83
S2	Screenshot of the gene AmTSAR1 (AH-019582) region	84

List of Tables

1	Tracks available in PopAmaranth	38
S1	Summary of non-redundant trees to which we found significant D-values. . .	75
S2	Demographic parameters estimated for different demographic scenarios using Fastsimcoal2.	76
S3	Crosses within and between species along with primer pairs used to validate the crosses.	77
S4	List of primers used to validate the crosses for the F1 plants	78
S5	List of accessions used in this the study. Names follow USDA germplasm ID	79
S1	List of samples evaluated.	85
S2	List of all tracks available on PopAmaranth at the time of publication. . .	90

Publications

Gonçalves-Dias, J.*, Singh, A.*, Graf, C., & Stetter, M. G. (2023). Genetic incompatibilities and evolutionary rescue by wild relatives shaped grain amaranth domestication. *Molecular Biology and Evolution*, 40(8), msad177.

Gonçalves-Dias, J., & Stetter, M. G. (2021). PopAmaranth: a population genetic genome browser for grain amaranths and their wild relatives. *G3*, 11(7), jkab103.

Authors Contributions

Chapter 2 JGD processed the whole-genome-sequencing data, performing gene-flow analysis per individual, per species and per region, represented in panel B of Figure 2 and Figures S3 and S4. JGD also performed the topology inferences displayed in Panel C of Figure 2. Further, JGD conducted the genome-wide local ancestry inference, calculating per site, per scaffold, per individual, and summarizing it per population for the respective plots on Figure 3 and S5. and conducted a genome-wide analysis of gene flow and local ancestry inferences. In addition, JGD calculated genome-wide the selective sweeps per species and calculated pairwise gene flow to detect possible overlaps as displayed in Figure 4. AS performed genetic load analysis and demographic modelling shown in Figures S6 and S2. CG performed selfing lines growth for multiple generations and the experimental crosses for incompatibility. CG performed the validation of the crosses. JGD and CG executed data collection and measurements, with the results summarized in Figure 6 and table S3. JGD, AS and CG prepared figures and tables. MGS, JGD and AS wrote the manuscript. All authors discussed the results and edited and approved the manuscript.

Chapter 3 JGD processed the data and conducted summary statistics calculations for each population. JGD created the Genome Browser, customized the layout, pre-processed the data and integrated it as tracks with extra summaries. MGS integrated the browser into amranthGDB. JGD prepared figures and tables. MGS and JGD wrote the manuscript. All authors discussed the results and edited and approved the manuscript.

JGD was also responsible for code upload and availability for the studies.

(**JGD** José Gonçalves-Dias, **AS** Akanshka Singh, **CG** Corbininan Graf, **MGS** Markus G. Stetter)

1 Introduction

1.1 Genomics and Computational Biology enable Population Genomics

Genomics comprises the study of genomes diversity, their structure, function, or regulation. Studying the proximity and distance between species helps researchers to depict their relationship and reconstruct their evolutionary history. Learning how those changes are reflected in today's genomes can help to create better models and predictions for the future (Ritland and Clegg, 1987). Since the unveiling of the DNA structure by Watson and Crick in 1953 and the assembly of the first complete genome using Sanger capillary sequencing in 1977, the field of genomics has rapidly expanded, enhancing its capabilities and resolution. The past few decades have seen significant advancements in technology, resulting in a huge leap in population genomics. Landmark projects like the release of the whole-genome of *Arabidopsis thaliana* (Initiative, 2000) and the Human Genome Project (Venter et al., 2001) made whole genomes accessible for researchers. Meanwhile, the output of DNA sequencing has increased, while the costs have decreased significantly (Wolinsky, 2007; Reuter et al., 2015). The combination of high-throughput parallel short-read sequencing technologies and more advanced computational tools has allowed for a finer-scale in research than ever before (Wu et al., 2010). Expanding reference panels such as the 1001 genomes project (Weigel and Mott, 2009) and whole-genome sequences opens a new window for comparative studies. While population studies date from decades (Charlesworth and Charlesworth, 2017), the convergence of high-throughput sequencing and computational capacity created new opportunities for unprecedented fine-grain understanding of the evolutionary processes leading to current populations. On the other hand, the growing amount and scale of generated data make scientists move beyond their individual disciplines (Eddy, 2005). Computational tools can address a new variety of research questions in the biological sciences, never possible before. Particularly, Computational Biology takes advantage of these technological leaps to analyze and interpret biological data. It encompasses a wide range of approaches, including bioinformatics, statistical analysis, and machine learning.

Plant population genomics typically involves high-throughput genomic technologies, such as DNA sequencing, generating large amounts of genetic data through computational methods to understand better plant species' evolution, adaptation, and diversity (De Wit et al., 2012).

Food security presents a major challenge for humanity as a society. With the increased necessity of food yield for an ever-growing population (Kumar and Bhalothia, 2020). The application of new knowledge on previously unexplored crops contributes to addressing the challenges in the current context

of a rapidly changing climate.

1.1.1 Resources in Computational Biology increase accessibility of Genomics results

The complexity of the growing amount of data, study systems, or downstream analyses can quickly result in the accumulation of haphazard information that is challenging to access, interpret, or reproduce. The advent of whole-genome data resources has two marked, complementary effects on the relatively new discipline of bioinformatics. Firstly, the flood of data creates a need and demand for new tools. Secondly, the unprecedented extent, diversity, and increasing completeness of the data sets are creating opportunities for new approaches to discovery based on computational methods (Denn and MacMullen, 2002). Further, these data are being generated simultaneously by many different groups spread all over the globe. Groups consist of diverse researchers, requiring information tailored to their specific contexts. Establishing resources that the scientific community can explore, contribute to, and access is necessary. The sharing of information allows not only the collective knowledge to grow but also a more efficient usage of resources, avoiding duplication of work and waste of high-cost and, many times, hard-to-obtain genomic information. Better accessibility allows for easier and clearer data exploration, increasing the potential for collaboration between dry and wet lab scientists. Such examples are databases such as Ensembl (Hubbard et al., 2002), NCBI (Pruitt et al., 2005) or Phytozome (Goodstein et al., 2012), which host archives for read sequencing in respective projects but also reference genomes or annotations. These references are vital for comparative studies and establish a common ground for further studies. The ubiquitous access to genomic data, primarily establishing standard file formats, pushes the development of resources and tools and builds the foundation for Computational Biologists to explore them in the constantly generated sequencing information.

The availability of genome-wide diversity data of crops and their wild relatives has already impacted the identification and study of candidate genes of agronomic significance, being it a loss for seed shattering, grain quality, or pericarp color, among others (Huang et al., 2012; Hufford et al., 2012; Wang et al., 2020a). Although many of the resources allowed the expansion of knowledge extraction, those tools are still very hard to access for non-computational scientists. To bridge the gap between dry and wet labs, new visualization resources are necessary to provide access to the summary statistics in a user-friendly way. Intuitive interfaces such as genome browsers provide graphical interfaces for analyzing, searching, and retrieving genomic sequences and annotation data. Genome browsers allow researchers to quickly understand patterns and identify genes or regions of interest in multi-omics or multi-species analysis. While genome browsers became widely spread for genetic studies, they are primarily present

for model species and mostly for animals or humans. Resources like Ensembl (Hubbard et al., 2002), UCSC (Karolchik et al., 2003), or TAIR (Swarbreck et al., 2007) are invaluable resources for the scientific community. Despite this recognition, this sort of tool is still scarce in population genomics, particularly in plant sciences.

1.2 Population Genomics to study the forces driving evolution

Genetic variation can be explained by a combination of selective pressures and demographic factors, such as population size and structure (Novo et al., 2022). Population structure alone can have a substantial effect on genetic differentiation. Previous studies have demonstrated that population structure can correlate with local environmental adaptation, such as molecular traits that provide high-temperature adaptation (Pradhan et al., 2016). The subdivision of populations can create heterogeneous effects along the genome between species (Lawson et al., 2012). An investigation into population structures using genomic data revealed that the distribution of genetic variation varies greatly across the genome. Some regions show high levels of genetic differentiation between populations, while others demonstrate little or no evidence of population structure (Lohmueller et al., 2009).

Organisms change over time in response to selective pressures in their environments (Crozier et al., 2008). This can lead to the development of new traits that are advantageous for survival and reproduction, ultimately resulting in the emergence of new species (Baack and Rieseberg, 2007). For example, if a population of organisms is isolated from other populations by a physical barrier (such as a mountain range), they may experience different selective pressures and evolve differently. If these differences become large enough, the two populations may become reproductively isolated, affecting their exposure to genetic drift differently. These changes in allele frequencies can accumulate enough to evolve into separate species eventually (Feder et al., 2012).

The effective population size (N_e) can also affect the rate of evolutionary change (Wright, 1931). Although effective population size can help determine the rate of change in the demographics and composition of a population (Charlesworth, 2009), other forces contribute to these alterations.

One of these forces is genetic drift, which introduces random changes in allele frequencies in a population. These changes in each generation can alter gene variant frequencies and ultimately expand rarer variants or make other traits disappear (Gossmann et al., 2011). Genetic drift can alter allele frequencies, causing rarer variants to expand and shift from the Hardy-Weinberg equilibrium. The equilibrium principle posits that the prevalence of different neutral alleles will remain consistent across generations in sizeable populations. Nonetheless, as time passes, stochastic changes in these frequencies between generations will likely result in the eradication of certain alleles and the fixation of others (Zeigler, 2014;

Lequime et al., 2016). The strength of genetic drift is strongly associated with population size (Rich et al., 1979). According to Sewall Wright's hypothesis, (Wright, 1931) drift can be more dominant in smaller populations as these populations are more susceptible to the accumulation of drift load (Cruzan, 2022).

The genetic diversity of a population is also heavily influenced by its size. Generally, larger populations tend to have a greater amount of genetic variation, whereas smaller populations are more vulnerable to a loss of diversity through genetic drift (Athrey et al., 2018). Conversely, smaller populations typically exhibit lower genetic diversity when compared to their larger counterparts. As a result, it is crucial to take population size into account when examining genetic variation (Montana et al., 2017). The presence of this standing genetic variation provides the genomic pool for a species to adapt to potential new constraints or develop new characteristics that can be advantageous (Masel, 2011; Burke et al., 2014). The maintenance, diminishing, or expansion of the strength of drift is influenced by selection. The two forces compete in the evolutionary dynamics of range expansions (Weinstein et al., 2017). When directional selection acts, it leaves signatures on the patterns of nucleotide polymorphisms, usually classified as selective sweeps (Stephan, 2019). These selective sweeps, the rise in frequency of certain alleles in a region of the genome, reduce the nucleotide diversity by increasing genetic differentiation between populations and deviating the allele frequency from neutral expectation. Exposure to different conditions can increase the selective pressure acting on the genomes. A particular case of selective pressure is the domestication of species that can affect genetic diversity (Hammer, 1984; Doebley et al., 2006). Repetitive human selection of desired traits leads to allele frequency change along the genome. Thus, domestication competes in an intricate balance with selection, genetic drift, and gene flow. While the first domestication tends to be sustained and directional, the other forces can confound these signals (Simon and Coop, 2023).

Gene flow is the transfer of genetic material between populations through hybridization, which can cause the incorporation of immigrant genomes via sexual reproduction or hybridization. In addition, the process of gene flow can have both positive and negative effects on genetic variation. In some cases, high gene flow can lead to the homogenization of genetic diversity among populations, limiting local adaptation and reducing the capacity to withstand environmental change (Tigano and Friesen, 2016). However, gene flow can also introduce novel genetic variation into populations, especially when coupled with genetic drift and mutation (Appiah-Madson et al., 2022). This can increase the overall genetic variation within a population, particularly in larger populations with higher levels of gene flow (Gompert et al., 2021). Additionally, gene flow can facilitate the spread of adaptive genetic variants across populations, allowing for local adaptation to occur in diverse habitats. The impact of gene flow on genetic variation is influenced by several factors, such as population size, migration rate, and the presence of selection pressure (Frantz et al., 2015). The effects of gene flow on a population depend on the magnitude and direction of the gene

flow, as well as the genetic characteristics of the population and the environment.

Overall, gene flow plays an essential role in sustaining and improving the genetic diversity and survival of a population. It can introduce new genetic diversity into a population that may have experienced a population bottleneck, thereby increasing the species' chances of survival. This process can also be the key to a faster recovery of the species after difficult conditions such as the glacial periods in Lower Guinea (Budde et al., 2013). Gene flow can also contribute to new characteristics, such as features for domesticated species or increased adaptation to a region. On the other hand, for well-established populations, the introgression of maladaptive genes can reduce the overall fitness and make gene flow a non-desirable mechanism (Ellstrand et al., 1999; Telschow et al., 2006). Gene flow can have different effects on different parts of the genome, depending on the genetic characteristics of the region and the selective pressures acting on it. For example, genes that play a role in reproductive fitness may be affected more by gene flow than genes that have little effect on reproductive success. Although some barriers to gene flow have been identified, many mechanisms that control the facilitation or restriction of gene flow are not yet fully understood. Ecological differentiation (Shapiro et al., 2012), genomic rearrangements (Rieseberg et al., 1995), and population density (Telschow et al., 2006) are among the factors that can influence gene flow. Further research is needed to elucidate the complex interactions underlying gene flow and its role in shaping genetic diversity and adaptation in populations.

By combining genomics and computational biology disciplines, we can better uncover the forces and patterns of genetic diversity and natural selection shaping populations. The application is particularly relevant for minor crops where fewer resources, studies, and knowledge are available. The combination of these two disciplines can make research more accessible in these species. Population genomics studies genetic variation within and among populations using genomic data (Luikart et al., 2019). One of the critical characteristics of population genomics is that different types of genetic variation are typically heterogeneously distributed along the genome (Causse et al., 2013). Instead, different types of genetic variation tend to be clustered in specific genomic regions, reducing its variation population-wide (Lam et al., 2010). Studying those regions and comparing patterns between populations can offer insights into their history. We can examine these markers in greater detail at the level of single nucleotide polymorphism (SNP) using available high-resolution data (Mitchell-Olds et al., 2007; Yang et al., 2011; Varshney et al., 2019). By analyzing the distribution and frequency of these markers, population genomics can provide insights into the history of a population, including information on when and where a population originated, how it has migrated over time, and how it has adapted to different environments. Particularly, studying the evolution of populations and the driving forces behind their current state can fill gaps in population history (Meyer et al., 2016). By performing variant calling, we can learn about the percentage of heterozygous phenotypes found in each species. Comparing the number of shared variants between

populations (genetic differentiation or F_{ST} (Wright, 1950)) can provide a proxy for proximity between populations. With SNP information, we can also learn about the diversity of populations. For example, with the Wu and Waterson estimator (Watterson, 1975), we can estimate the population mutation rate deviation from the population's observed nucleotide diversity. Tajima's D (Tajima, 1989) can give an indication of possible acting selection, population size changes and distinguishing between random and non-random evolution.

1.2.1 Domestication as framework for evolution

The domestication of animals and plants, which facilitated the transition from hunting and gathering to farming, marks a significant milestone in the evolution of humans. The transformation enabled humans to settle down and form communities, laying the foundation for the complex societies that exist today (Purugganan, 2019; Stetter, 2020). Domestication can be seen as a coevolutionary process between humans and domesticated species, involving specialized mutualism where humans control the fitness of the species to gain resources or services (Purugganan, 2022). In addition to the other competing forces, domestication creates extreme selective pressure for preferred traits (Innan and Kim, 2004). Crops suffer from the so-called "domestication syndrome" that involves similar phenotypic changes that strongly differentiate them from their wild ancestors (Stetter, 2020). However, domestication is a progressive process with different degrees. Many crops have not acquired all the ideal agroecological traits that would make them "fully domesticate" (Meyer et al., 2012). Some of these traits include size, color, and other characteristics that favor culture and harvesting, such as resistance to drought or loss of seed shattering (Ross-Ibarra et al., 2007; Purugganan and Fuller, 2009; Abbo et al., 2014). Some plant species are "incomplete domesticates," lacking some of those traits despite their historical cultivation. By studying the genomes of such species, we can learn about the genetic background that leads to the genomes of current populations. Given their close relationship, learning about domestication is not only about a species' history but also about human history as a society. Domestication typically affects population demographics (Gaut et al., 2018). As the first species after hybridization are not yet completely adapted, it tends to shrink the population size, leading to a bottleneck (Eyre-Walker et al., 1998). Genetic diversity tends to decrease as variants of genes that translated into favorable phenotypes increase in frequency and variants that confer wild characteristics decrease (Doebley, 2006). These variations leave signatures that can be used for comparative studies to understand the genomic landscape of domestication. Further, the impact of gene flow on plant domestication is substantial. The introgression from wild plants into newly domesticated populations can facilitate adaptation to different environments (Moreno-Letelier et al., 2020) and minimize the accumulation of harmful mutations in smaller populations. Moreover, gene flow

from relatives with larger population sizes can alleviate genetic load in populations with smaller effective population sizes and may even rescue a small population from evolutionary decline. Nonetheless, it is worth noting that crop-wild hybrids are not typically expected to excel as crops, nor fit for the wilderness (Stetter et al., 2020).

Domestication can also be a driver of speciation (Hilton and Gaut, 1998). The strong selective pressures on specific traits lead to the progressive formation of reproductive barriers, contributing to the isolation of previous wild populations and creating new species. However, other cases of continuous exchanges exist between the "newly formed" domesticated species and their wild counterparts (Tenailon et al., 2023). In the case of domesticated populations, gene flow might play a pivotal role. For example, during expansion to new habitats, the contribution of beneficial alleles from locally adapted populations can improve the fitness of the crop (Janzen et al., 2019). In other cases, it can also result in "feralized" species, where the crop acquires wild characteristics, creating a new intermediary species (Gering et al., 2019; Wu et al., 2021). Understanding the genetic consequences of the evolution of wild and domesticated crops and the balance between all these forces can provide insights into the evolution of populations and the effects of domestication as a speciation catalyst. Comprehending such mechanisms can be helpful in enhancing future breeding programs by leveraging the genomic insights obtained from such studies.

1.3 Revolution in non-model plant research

The development, maintenance and extension of biological resources can be laborious and expensive; therefore, the usage of model organisms is essential for life science research. Model organisms have several characteristics that facilitate experimental manipulation and analysis, such as small size, rapid growth, short generation time, and amenability to genetic transformation and recombination (Cesarino et al., 2020). Model organisms also benefit from wide usage in multiple research facilities, allowing a continuous extension of its resources, from genome sequences, gene annotation, or experimental validation (Armengaud et al., 2014). While convenient, the major model organisms are not necessarily the best possible system for all the questions. The advent of modern tools, the lowering costs for downstream sequencing, and the increasing computational capacity for model generation and data analysis allow a broadening of research studies to less-known systems (Russell et al., 2017). Expanding studies to multiple organisms, such as plants, grasses, or invasive species, will allow us to comprehend these populations' evolutionary histories better and how they intertwine. Research on non-model organisms provides the opportunity to explore new and previously uncharted avenues of discovery (Russell et al., 2017). Studies in non-model species have already revealed mechanisms previously unknown, being cancer resistance in elephants and other long-lived mammals (Seluanov et al., 2018) or senescence mechanisms in plants

(Bernard et al., 2020; Wang et al., 2020b). By choosing plants as the object of study, we can benefit from a combination of advantages such as lower costs, higher throughput, and shorter generation times. Besides, studying crops can provide us with insights into both the species and the human populations that have cultivated them.

1.4 Amaranth as a model system for studying genomic mosaic of domestication

The *Amaranthus* genus comprises over 70 species, with ecological varied effects, existing in the wild, as weed, or as a domesticated plant (also referred to as grain amaranth). *Amaranthus* belongs to the dicotyledonous *Amarantaceae* plant family and is divided into three subgenera, *Acnida*, *Amaranthus* and *Albersia* (Costea et al., 2001). It can grow in width and height up to 4 meters high (Pastor and Acanski, 2018). Amaranth has historically served multiple purposes, with edible seeds and leaves, as an ornament (Sauer, 1950), or as part of religious ceremonies and traditions (Sauer, 1993). Its gluten-free seeds contain a high lysine, fiber, and protein content and are low in saturated fats (Aswal et al., 2016; Joshi et al., 2018). Amaranth is a versatile grain that can be ground into flour, popped like popcorn, or flaked like oatmeal. It is already commercially available in these variable forms, albeit on a small scale.

While it can be found widely around the globe, grain amaranth originated from the Americas. Amaranth has a long history of cultivation and use by humans, going back to around 8,000 years ago. It has been exported, cultivated, and used as a staple food and medicinal plant in many parts of the world. Amaranth possesses some unique characteristics, which include the ability to tolerate drought and other environmental stressors, its high nutrient content and medicinal properties, and its potential to be used in sustainable agriculture systems (Sauer, 1950, 1967a; Stetter et al., 2020). Further, as a C4 plant, having a relatively small genome (Stetter et al., 2017a; Lightfoot et al., 2017a), and rapid breeding capability (Stetter et al., 2016) makes amaranth a great model for crop evolution and domestication studies (Stetter et al., 2020).

Despite its long cultivation history in the proximity of other fully domesticated plants such as maize (Stitzer and Ross-Ibarra, 2018), tomato (Razifard et al., 2020), or common beans (Rendón-Anaya et al., 2017), grain amaranth is still incompletely domesticated. The co-localization with the different species indicates that alone, the lack of selection, being it conscious or unconscious (Purugganan, 2019; Yang et al., 2019) is not enough to explain this incompleteness of domestication. Lack of standing genetic variation, trait pleiotropy, polygenicity, network integration, accumulation of genetic load, admixture, and gene flow can influence the path of domestication (Stetter, 2020; Gonçalves-Dias et al., 2023).

Further, wild and domesticated amaranth can intercross, making the definition of its populations

more complex. The continuous exchanges between amaranth populations can allow gene flow to provide adaptive alleles or to act as a genetic rescuer, countering the loss of diversity during the bottleneck occurring during domestication and mitigating crop vulnerability (Tenailon et al., 2023). Contrastingly, gene flow can also allow the domesticate to return to the "wild" and be viable without human intervention again (Gering et al., 2019).

The taxonomic classification of amaranth has long been a subject of debate due in part to the plasticity of the genus and the difficulty in distinguishing between grain and non-grain species (Sauer, 1967a; Kietlinski et al., 2014). Further, the domestication history of these populations has also been the subject of intense discussion (Sauer, 1967a; Kietlinski et al., 2014; Stetter et al., 2017a).

The release of reference genomes for the grain amaranth species *A. hypochondriacus* (Lightfoot et al., 2017a) and *A. cruentus* (Ma et al., 2021) paves the way to high-resolution and comparative genomic studies. A recent study, which sequenced more than 100 amaranth individuals, suggested the independent domestication of the three grain amaranths from a single *A. hybridus* species, twice in Central America and a third time in South America (Stetter et al., 2020). This was supported by a higher genetic differentiation between crop species than between crops and their wild ancestors. Nucleotide diversity was also lower on the crops than on the wild ancestors (Stetter et al., 2020; Gonçalves-Dias and Stetter, 2021), expected from populations that went through domestication bottlenecks (Hufford et al., 2012). Despite these differences, amaranth populations are still weakly differentiated and mostly able to hybridize between themselves, even when strongly geographically segregated. This hybridization gives the opportunity for gene flow to occur inter and intra wild and domesticated populations. While its presence has been reported Stetter et al. (2020), gene flow still needs to be studied in fine detail for grain amaranth.

Improving genomic knowledge and breeding techniques brings new opportunities for underutilized and underexplored alternative crops such as amaranth. Understanding the evolution of wild crop systems and the genomic landscape can prove an invaluable resource for a more capable, diverse, and sustainable agriculture with higher nutritional quality and better climate resilience.

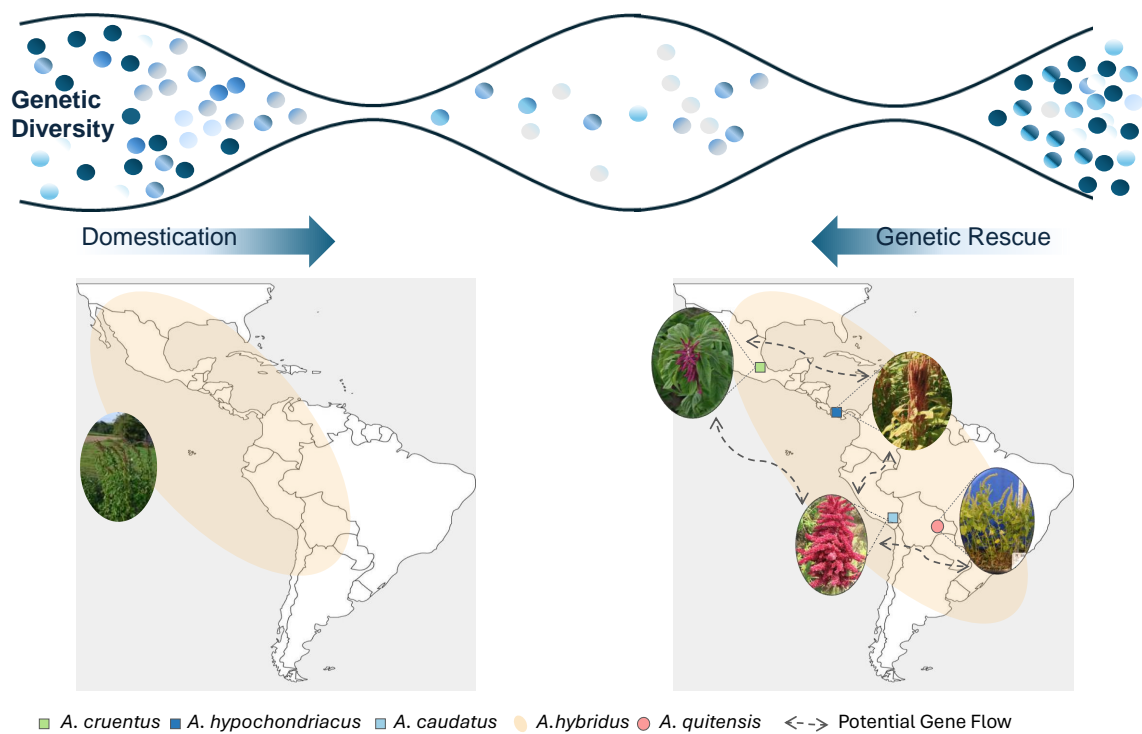


Figure 1: During the process of domestication, a bottleneck in population size occurs, reducing the genetic diversity in the population. The exchange of genetic material between populations, including wild populations, can help in the resilience of those domesticated populations (Genetic Rescue). At the bottom are illustrated the different amaranth populations included in this thesis and their geographic distributions. The possibility and magnitude of their exchanges are evaluated in detail.

1.5 Aims of the Thesis

The main goal of the present thesis is to investigate the post-domestication contact between populations, particularly gene flow and their effect on their genomes, making population genomic resources of amaranth available to the research community. While the evolution of domesticated species, particularly crops, has been perceived as a single linear process from a wild ancestor to its domesticated species, recent studies have demonstrated that this process is more complex for some species.

Here, I apply genomic data from whole-genomes of the crop-wild system of amaranth that originated from Central and South America to depict its evolutionary history by studying its current populations' genomes. Particularly, I focused on five sub-species of the *Amaranthus* genus, two wild (*A. hybridus* and *A. quitensis*) and three domesticated (*A. cruentus*, *A. hypochondriacus*, and *A. caudatus*). *A. cruentus* and *A. hypochondriacus* can be mostly found in Central America, whilst *A. caudatus* and *A. quitensis*

are native to South America. *A. hybridus* is spread all over the Americas (Figure 1).

In Chapter 2, the amount and distribution of gene flow and genetic load were quantified along the genome of the previously described amaranth species. Furthermore, we assess experimentally putative genetic incompatibilities between the different amaranth species.

In Chapter 3, I aim to provide an accessible representation of the genetic variation of amaranth populations during domestication and convergence across crops. The new resource, PopAmaranth, intends to make this research available and explorable for the interdisciplinary scientific community.

NB: Chapters 2 and 3 are transcribed *ipsis verbis* of their corresponding publications.

2 Genetic incompatibilities and evolutionary rescue by wild relatives shaped grain amaranth domestication

Abstract

Crop domestication and the subsequent expansion of crops have long been thought of as a linear process from a wild ancestor to a domesticate. However, evidence of gene flow from locally adapted wild relatives that provided adaptive alleles into crops has been identified in multiple species. Yet, little is known about the evolutionary consequences of gene flow during domestication and the interaction of gene flow and genetic load in crop populations. We study the pseudo-cereal grain amaranth that has been domesticated three times in different geographic regions of the Americas. We quantify the amount and distribution of gene flow and genetic load along the genome of the three grain amaranth species and their two wild relatives. Our results show ample gene flow between crop species and between crops and their wild relatives. Gene flow from wild relatives decreased genetic load in the three crop species. This suggests that wild relatives could provide evolutionary rescue by replacing deleterious alleles in crops. We assess experimental hybrids between the three crop species and found genetic incompatibilities between one Central American grain amaranth and the other two crop species. These incompatibilities might have created recent reproductive barriers and maintained species integrity today. Together, our results show that gene flow played an important role in the domestication and expansion of grain amaranth, despite genetic species barriers. The domestication of plants was likely not linear and created a genomic mosaic by multiple contributors with varying fitness effects for today's crops.

2.1 Introduction

Evolution and speciation has long been viewed as a linear process with a single ancestor giving rise to one or more derived species. Genomic data of large samples have revealed ancestry of different species within modern populations in a number of species (Harris and Nielsen, 2016; Niu et al., 2019; Chomicki et al.,

2020; Kozak et al., 2021; Orlando et al., 2021; Lv et al., 2022). Particularly in plants where reproductive barriers are often weak, the potential for the exchange of genetic material between related species is rather high (Ostevik et al., 2016; Osuna-Mascaró et al., 2023; Sheidai and Koohdar, 2023). Yet, the observation of exchanging genetic material between species is often seen as an exception and as a minor contribution to the genomic composition of a species.

Gene flow describes the process of exchanging genetic information between populations or even species (Rieseberg and Burke, 2001). While interbreeding between populations might be frequent, the manifestation of gene flow between locally adapted populations or even species is thought to be rare as it would decrease fitness. Nevertheless, beneficial gene flow has been shown to be an important source of variation for local adaptation (Crispo, 2008; Sexton et al., 2011; Ellstrand, 2014; Tigano and Friesen, 2016; López-Goldar and Agrawal, 2021). At least partial fertility of hybrids is required, and compatibility between donor and recipient determines the intensity of gene flow (Aguillon et al., 2022). In the course of speciation, the ability to form viable hybrids can be lost, and reproductive barriers that prevent gene flow can evolve. Hence, hybrid incompatibility can hinder gene flow and lead to reproductive isolation. Yet, reproductive isolation in plants is often incomplete or is circumvented by intermediate populations, allowing for gene flow even between different species.

The domestication of crops and animals led to a major transition in human lifestyle and had a profound impact on the genetic makeup of the domesticates (Doebley, 2006). Crop domestication can be seen as rapid evolution, often leading to speciation. Even more than speciation in general, crop domestication has long been described as a linear process starting from one wild species evolving through strong directional selection into a domesticate. However, this view has been challenged in recent years, where gene flow from wild relatives have been documented in a number of crops, including maize (Ross-Ibarra et al., 2009), rice (Yang et al., 2012), barley (Civán et al., 2021), sorghum (Sagnard et al., 2011), tomato (Razifard et al., 2020), *Brassica* (Saban et al., 2023) and others (Luo et al., 2007; Ding et al., 2022; Page et al., 2019; Liu et al., 2019). Reproductive isolation of the crop from its wild relatives would ensure the maintenance of domestication traits, hence, the success of domestication (Dempewolf et al., 2012). Gene flow from wild relatives would have led to a reduction of domestication-related phenotypic changes. Early generations of crop-wild hybrids would be rather unfit as wild plants or crops, as their adaptive traits strongly differ (Janzen et al., 2019; Stetter, 2020). Yet, gene flow from wild relatives could have increased the genetic variation in early crops, which could have been beneficial to increase adaptive potential (Smith et al., 2019). In addition, gene flow from locally adapted wild relatives has been shown to have provided alleles that allowed the crop population to establish in novel environments (Van Heerwaarden et al., 2011; Hufford et al., 2013; Wang et al., 2021).

Plant domestication has likely been driven by demographic changes and directional selection. Domes-

tication bottlenecks reduced the effective population size, leading to an accumulation of mildly deleterious alleles in the population (Gaut et al., 2018). Selection on major effect domestication genes might have allowed hitchhiking of linked mildly deleterious alleles (Sedivy et al., 2017). Together these effects increased genetic load – the accumulation of deleterious alleles – in the crop population (Bertorelle et al., 2022). Several studies have assessed the accumulation of genetic load in domesticates compared to their wild relatives, e.g., rice (Lu et al., 2006; Xu et al., 2006; Nabholz et al., 2014), maize (Rodgers-Melnick et al., 2015; Gaut et al., 2015), and soybean (Kim et al., 2021). While an accumulation of genetic load has been detected in many domesticated species, no increase has been detected in sorghum, potentially due to the transition to selfing in the crop (Lozano et al., 2021). Gene flow from wild relatives with larger effective population size into crop populations might have reduced genetic load in crops (Stetter, 2020). In sorghum, gene flow between early landraces resulted in decreased genetic load across landraces, but no variation in genetic load was observed among landraces with or without gene flow (Smith et al., 2019; Lozano et al., 2021). However, such evolutionary rescue by gene flow from wild relatives into domesticates has received little attention. We study the effects of gene flow on genetic load in a three times domesticated crop and its wild relatives to understand the evolutionary role of gene flow and genetic load during crop domestication.

Grain amaranth is a nutritious pseudo-cereal from the Americas that has been domesticated three times from one ancestral species (*A. hybridus*). Two grain amaranths were domesticated in Central America (*A. cruentus*, *A. hypochondriacus*) and one in South America (*A. caudatus*) (Sauer, 1967b). The taxonomic complexity of the *Amaranthus* genus led to different domestication scenarios for the crop (Sauer, 1967b; Kietlinski et al., 2014; Stetter et al., 2017b). Population genetic and genome-wide selection signals suggest subpopulations of *A. hybridus* as ancestors for all three grain amaranths (Stetter et al., 2020, Figure 2A and S1). In South America, the closely related *A. quitensis* was potentially involved in the domestication process of *A. caudatus* and genome-wide signals of gene flow between grain amaranths and their wild relatives have been detected previously (Kietlinski et al., 2014; Stetter et al., 2020). While the three grain amaranth species have been cultivated as a crop in different regions of America for thousands of years, all three lack key domestication traits (Stetter et al., 2017b). A potential reason for the lack of domestication traits might be continuous gene flow from wild relatives that prevented domestication traits to fix (Stetter, 2020). Understanding the underlying genomic signatures of gene flow and selection could improve the understanding of the evolutionary history of crops (Meyer et al., 2012).

In this study, we use population-wide whole genome sequencing data and reveal the mosaic signature of gene flow along the genome of domesticated amaranths. We found strong signals of gene flow even between the three geographically isolated domesticated species. Besides post-domestication gene flow between crops, we also observed high levels of genetic exchange between the South American wild

species and the local crop. A hybridization experiment between three crops indicates genetic incompatibility between *A. cruentus* and the other two grain amaranths. This reproductive barrier might be a contributing factor to the different rates of gene flow observed between species despite their geographic distances. Gene flow from the wild ancestor *A. hybridus* into the domesticated amaranths reduced genetic load in the crops, but only a few positively selected regions were exchanged through gene flow between crop species. Gene flow might be an important source of genetic variation for crops, not only to provide adaptive alleles but also to reduce genetic load and allow further selection.

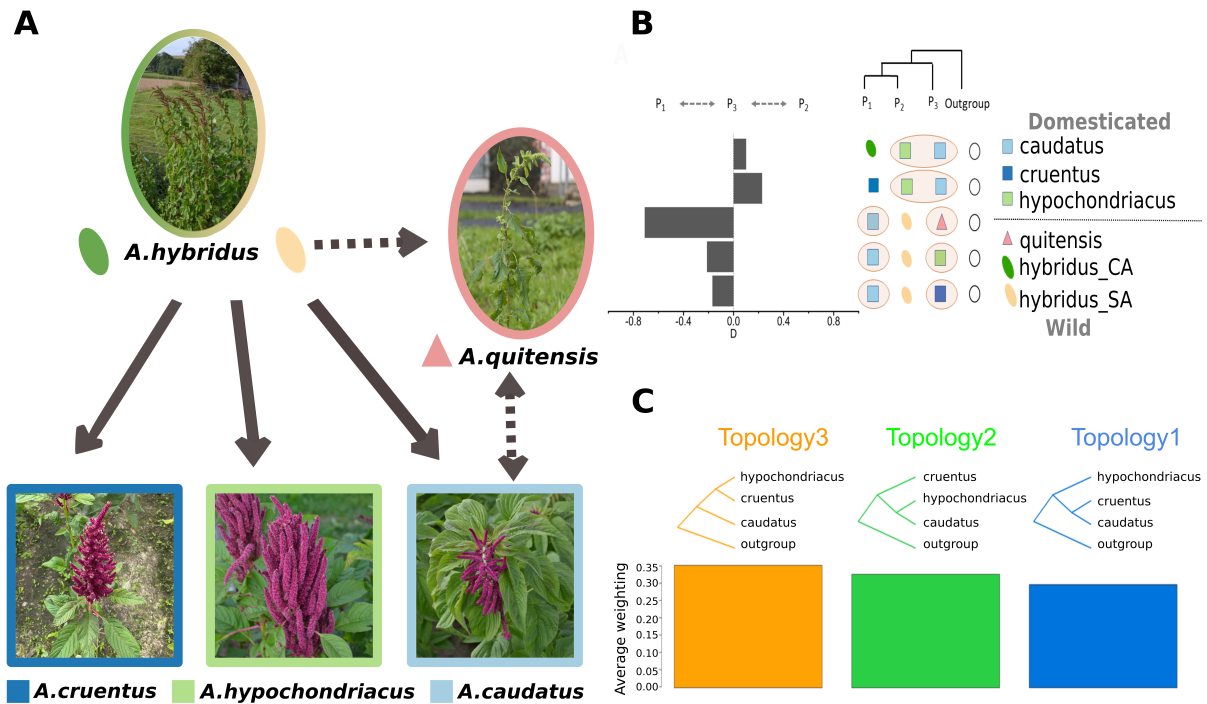


Figure 2: Genome-wide signals of gene flow. A) Schematic history of amaranth domestication. Amaranth has likely been domesticated three times independently from different subpopulations of *A. hybridus* (Central America (hybridus_CA) and South America (hybridus_SA)). *A. quitensis* is speculated to be an intermediate population between *A. hybridus_SA* and *A. caudatus*. Colors are consistent with the legend in B. B) Gene flow between amaranth populations (exchanging pairs highlighted in yellow). The D-value indicates the strength of gene flow. Only significant signals of gene flow are shown. C) Genome-wide summary of tree topologies along the genome inferred by Twisst. The proportion of each of the three topologies observed along the genome is shown in bars.

2.2 Results

Strong post-domestication exchange between crops and between crops and their wild relatives

Gene flow likely played an important role during the domestication of different crops (Janzen et al., 2019). In grain amaranth, genome-wide signals of gene flow between species have been previously reported (Stetter et al., 2017b, 2020). In order to quantify gene flow among different domesticated and wild populations, we measured ancient admixture using the D-statistic (ABBA-BABA) for all possible tree topologies using ANGSD (Korneliussen et al., 2014b) in whole genome sequencing data of six population samples of domesticated (caudatus: *A. caudatus*; cruentus: *A. cruentus*; hypochondriacus: *A. hypochondriacus*) and wild amaranth (hybridus_CA: *A. hybridus* from Central America; hybridus_SA: *A. hybridus* from South America and quitensis: *A. quitensis*). We identified gene flow between crop species and between crops and their wild relatives. We found ample gene flow even between geographically distant crop species (Figure 2B). The strongest signal was identified for the Central American crop hypochondriacus and the South American caudatus. This signal was robust even when changing the third species in the test (Figure 2B and Table S1). The test also identified gene flow between the South American crop species caudatus and the second Central American grain amaranth cruentus. However, the strength of gene flow between them was lower. This is also shown when the three grain species were tested in the same tree, where a significant level of gene flow between caudatus and hypochondriacus was identified ($D=0.22$; Figure 2B, second row), showing that the signal of gene flow between the two species is higher than the shared variation of the three crops. As both Central American grain amaranths were domesticated from Central American hybridus and the exact subpopulations of the ancestor that gave rise to each crop remain unknown, we could not test directly for gene flow between the two Central American grain species. The high level of exchange of genetic material between species was also shown by tree topology tests using Twisst (Martin and Van Belleghem, 2017). While the expected tree topology along the genome, with the two Central American crops being closest, was the most common, the other two alternative topologies were only slightly less abundant (Figure 2C). Altogether, more gene flow was observed between two allopatric crop species, caudatus and hypochondriacus, less gene flow was observed with cruentus.

We tested whether isolation with gene flow between hypochondriacus and caudatus was a better fitting model than a simple split without further gene flow by simulating demographic histories using Fastsimcoal2 (Excoffier et al., 2013). We investigated three alternative models: population split without gene flow, population split with one-time gene flow and population split with continuous gene flow (Figure

S2). The model for a population split with continuous gene flow was obtained as the best model (Table S2), suggesting that this is the most suitable scenario (Figure S2). In addition, this model predicts population splits estimates that coincide with previous observations (Stetter et al., 2020).

In addition to gene flow between crop species, we examined gene flow signals between the crops and their wild relatives. We found the strongest signal of gene flow between the sympatric South American grain crop *caudatus* and the local wild *quitensis* with a D-value of -0.71. When testing a scenario where we assumed *quitensis* as the ancestor of *caudatus*, which had been suggested previously (Sauer, 1967b), significant gene flow from South American hybridus was detected (Figures S3 and S4), indicating the intermediate role of *A. quitensis* between wild and domesticated species.

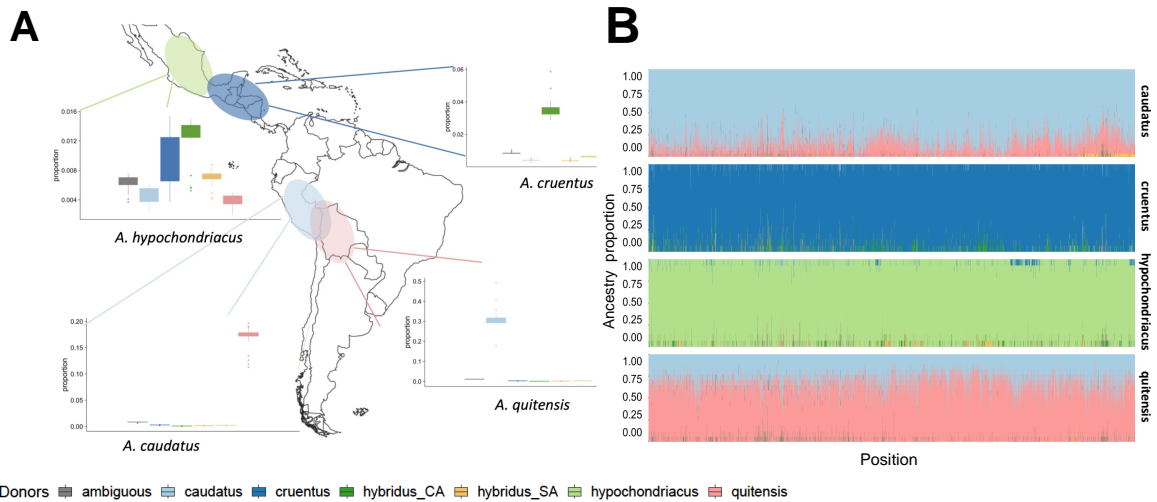


Figure 3: Variable ancestry across individuals and along the genome. A) Contribution of donor populations to individual recipients. Values within boxplots represent contributions by different donor populations to each individual of the recipient population. The Y-axis scale differs between plots. The schematic geographic range of populations. B) Population scale ancestry proportion along genomic positions of Scaffold 4. The proportion of the most likely donor population at a given SNP across all individuals in the recipient population. Each plot represents a recipient population, and colors represent donor populations. Exemplary scaffold, all scaffolds in Figure S5. Donor colors according to Figure 1.

2.2.1 Fine-scale gene flow reveals diverse local ancestry of grain amaranths

The genome-wide and population-wide gene flow analysis already showed the complex pattern of exchange of genetic material between the *Amaranthus* species. We found evidence of gene flow for distant and closely related species. To understand the species complex as a whole, we inferred the local ancestry (Lawson et al., 2012) along the genome of each individual using finestructure v4.1 (Lawson et al., 2012) and summarized them by population (Figure 3). We observed that the Central American grain species *cruentus* and *hypochondriacus* had less admixed backgrounds but shared ancestry tracks depending on the different individuals. Both had the highest donated portion from the wild *hybridus_CA* (Figure

3). *Hypochondriacus* had a heterogeneous contribution from the other species to different individuals (although with proportions less than 0.016 total), while *cruentus* had a more homogeneous contribution among individuals in the population. In South America, the overall pattern observed with population-wide introgression tests was also confirmed by the individual-based test (Figures S3 and S4). The South American populations *caudatus* and *quitensis* shared large amounts of ancestry, but this strongly varied between the individuals, between 10.8% and 18.5% of *quitensis* ancestry in *caudatus* individuals (Figure 3A), which also varied along the genome (Figure 3B).

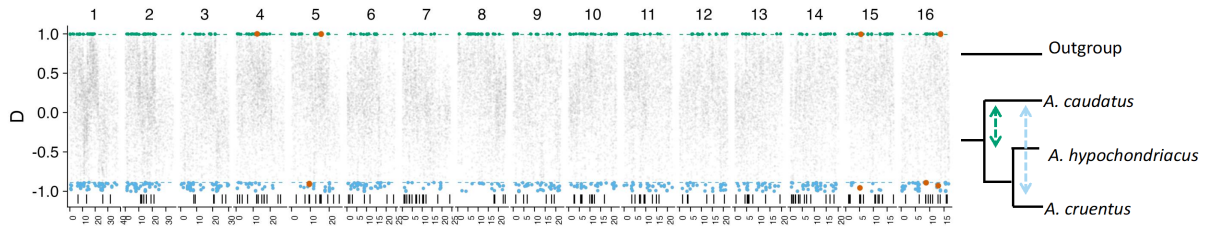


Figure 4: Fine-scale gene flow along the genome between domesticated amaranth populations. We used the D-value to calculate gene flow along the genome in 1000 SNPs windows. Each dot represents the D-value for a tree comparing *A. caudatus* with *A. hypochondriacus* or *A. cruentus*. Positive values are indicative of gene flow between *A. caudatus* and *A. hypochondriacus* and negative values between *A. caudatus* and *A. cruentus*. The top 0.1% of windows in each direction were colored. The bars at the bottom indicate previously detected selective sweep regions detected in *A. caudatus*. Orange dots denote overlaps between selective sweep and top gene flow signal.

We further wanted to understand whether gene flow is variable along the genome. Comparing ancestry signals along the genome of the different species showed that even the same region can have multiple donors in a species (Figure 3B and Figure S5). Using D-value in genomic windows, we scanned the genomes for gene flow signals in the trees with significant genome-wide signals. The previous comparison between crops showed the presence of gene flow between *hypochondriacus* and *caudatus* and between *cruentus* and *caudatus* (Figure 2). In the local scan, we observed similar signals as for global gene flow analysis; stronger gene flow between *caudatus* and *hypochondriacus* ($D > 0.992$ for top 1% windows) than between *caudatus* and *cruentus* ($D < -0.887$ for bottom 1% windows) (Figure 4). Windows representing significant gene flow from *cruentus* or *hypochondriacus* with *caudatus* did not overlap, suggesting that gene flow occurred independently between species.

To understand the potential reason for the observed high levels of gene flow between grain amaranth species, we combined gene flow signals along the genome with selection scan results (Gonçalves-Dias and Stetter, 2021). Despite the genome-wide distribution of gene flow between crop species, only a few selective sweeps overlapped with outlier windows of gene flow between *caudatus* and the other two crop species. We found 13 overlapping windows in total, eight in regions of gene flow between *hypochondriacus* and

caudatus and five between caudatus and cruentus. Despite the relatively low total number, the overlap was higher than expected by chance ($p=0.02$), suggesting beneficial gene flow between geographically distant crop relatives.

2.2.2 Introgression from wild ancestor mitigates increased genetic load in domesticated grain amaranths

In many crop species, a reduction in overall genetic diversity between wild relatives and the crop has been observed. This has been associated with population bottlenecks and directional selection during domestication (Gaut et al., 2018). Increased genetic drift and hitchhiking with selected alleles can lead to a higher genetic load in the domesticated species (Lu et al., 2006; Wang et al., 2017). We calculated GERP scores for the *A. hypochondriacus* reference genome from whole genome alignments with 15 diverse plant species of different relatedness as a proxy for deleterious alleles (Figure S6). We observed that two domesticated species (caudatus and cruentus) had a significantly higher total genetic load than their wild ancestor hybridus. The third crop species, hypochondriacus, had a higher load than hybridus_SA but lower than hybridus_CA (Figure 5A). This pattern remained even when using the *A. cruentus* reference genome (Ma et al., 2021) to calculate GERP scores, showing that the difference is likely not the result of reference bias when calculating GERP scores (Figure S7). The South American relative quitensis showed as high total load as the domesticated species. Partitioning of total genetic load from fixed and segregating sites showed that the domesticated species had a high fixed load but a lower segregating load than their wild ancestor (Figure 5B and C). Quitensis also showed high fixed load and low segregating load, in agreement with the small effective population size and low genetic diversity documented previously (Stetter et al., 2020).

To investigate if hitchhiking of deleterious alleles with selected loci led to the overall increase in genetic load in the domesticated species, we compared genetic load within selective sweep regions (Gonçalves-Dias and Stetter, 2021) with that of random non-sweep regions. We observed that the mean load per site in sweep regions was significantly lower than in control regions for all domesticates (Figure 5D). However, the mean GERP score per deleterious site in the sweep regions showed higher values than deleterious sites in control regions, suggesting a hitchhiking effect (Figure 5E). This suggests that strongly deleterious alleles might accumulate within selective sweep regions due to hitchhiking, but mildly deleterious alleles fix or increase in frequency due to increased genetic drift.

The abundant gene flow between grain amaranths and their wild relatives is expected to impact the patterns of genetic load. To evaluate the potential effect of gene flow on the accumulation of deleterious alleles, we measured the total load accumulated within introgressed regions from other species into a

recipient species as the proportion of introgressed load received per individual. The expected value if introgressed regions carry the same amount of load as random regions would be one, while the value is higher than one for deleterious introgression and less than one for beneficial introgression. The analysis showed that introgressed regions from *A. hybridus* into the domesticated reduced genetic load, while introgression between crops donated higher load in the recipient population (Figure 5F). Introgression from domesticated donors into the wild species resulted in a higher genetic load. These results might suggest that gene flow from populations with higher effective population sizes (wild relatives) could provide evolutionary rescue for the smaller populations (crops).

2.2.3 Hybrid incompatibilities between grain amaranths

All gene flow begins with the inter-mating between individuals of different populations. To further understand the process of gene flow and differences in observed levels of gene flow, we crossed multiple inbred lines of the three crop species. All three amaranth species are mostly selfing, but outcrossing is possible and occurs in the field. We selected three accessions from each crop species and crossed them within and between species, including selfings. All crosses produced viable F₁ seeds. Selfings and intra-specific F₁ plants grew without complications and set seeds (Figure 6, Table S3). The inter-specific F₁ plants of the 5 combinations between *A. caudatus* and *A. hypochondriacus* developed healthy and fertile plants. For crosses between *A. cruentus* and *A. hypochondriacus*, one of the three combinations led to a lethal phenotype, with unhealthy seedlings that did not survive the juvenile stage (Figure 6). Yet, both parents of this cross produced healthy and fertile offspring with other crossing partners, suggesting that incompatibility is the result of a specific allelic combination rather than an individual unfit parent (Table S3). All of the six combinations between *A. caudatus* and *A. cruentus* resulted in the lethal phenotype, suggesting a genetic incompatibility between these species. The difference in hybrid compatibility between grain amaranth species likely contributed to the observed patterns of gene flow.

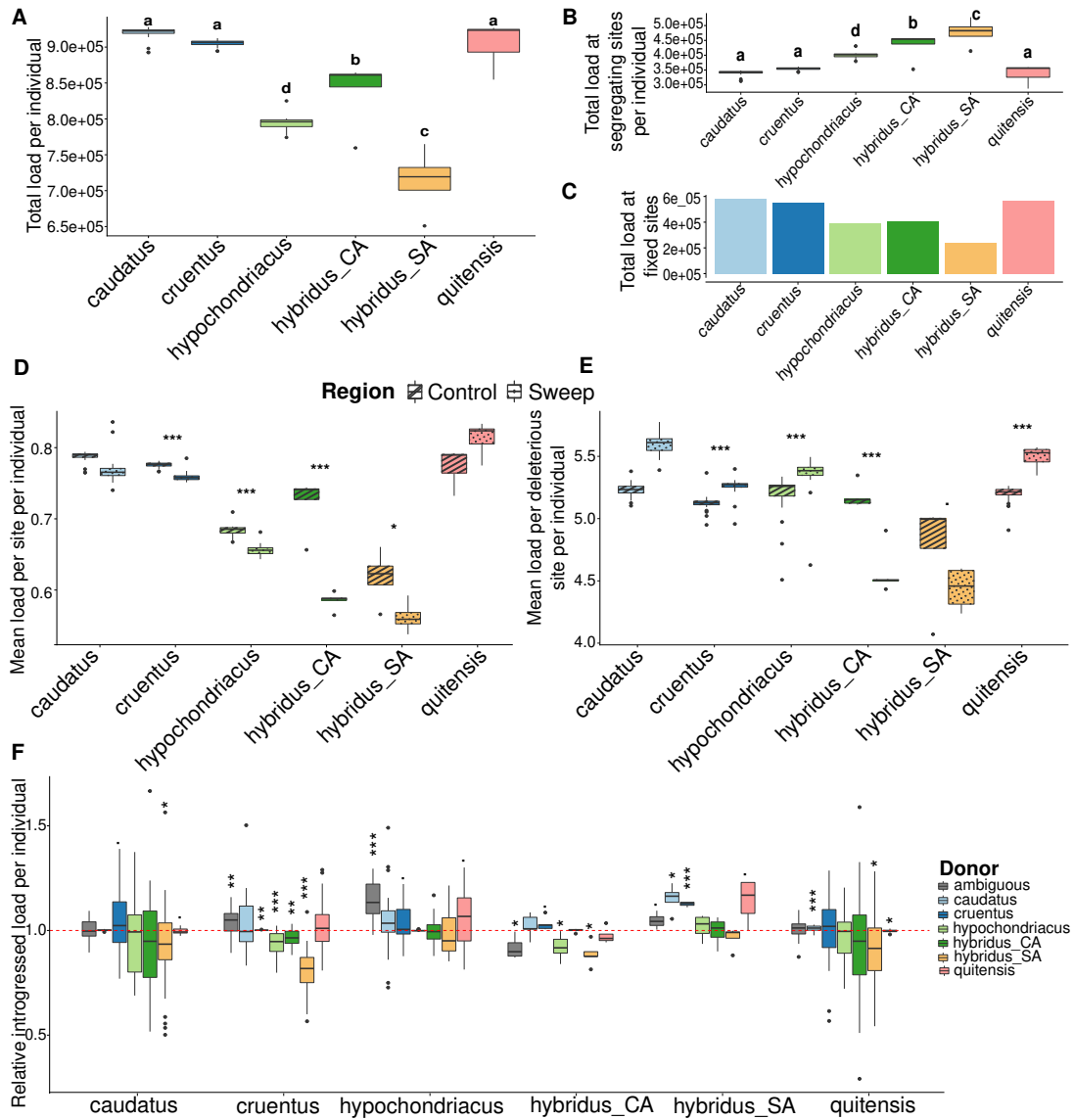


Figure 5: Genetic load in domesticated and wild amaranth. A-C) Genetic load was calculated as the sum of GERP scores for the derived allele per individual. A) Total genetic load per individual in domesticated and wild populations. Different letters above the box plot represent significant differences. B) Segregating genetic load in each population per individual and C) Fixed genetic load within each population. D) and E) Accumulation of genetic load in selective sweep regions and control regions (rest of the genome). D) Mean load of sweep/non-sweep region (including all sites in region); E) Mean effect (GERP score) of deleterious allele in region (only deleterious sites). Asterisks above box plot represent the significance level. F) Relative introgressed genetic load; load in introgressed regions relative to introgression received from the donor. A value greater than one represents increased load through introgression, while a value lower than one shows a reduction in load through introgression compared to the amount of introgression by the donor. A value of 1 shows, the expectation of equal load and introgression proportion (denoted by red dotted line). The asterisks above the box plot represent the significance level for one-sample t-test. (* - p-value < 0.05, ** - Pvalue < 0.01, *** - Pvalue < 0.001)




	<i>cau</i>	<i>cru</i>	<i>hyp</i>
<i>cau</i>	0/3		
<i>cru</i>	6/6	0/3	
<i>hyp</i>	0/5	1/3	0/4

Figure 6: Proportion of lethal F₁ phenotypes. The lower triangle shows the number of lethal combinations between accessions out of the total number of combinations. We considered a cross as "lethal" when all F₁ seedlings died within 20 days after planting. The upper triangle shows example images of phenotypes of inter-specific combinations. *cau*: *A. caudatus*, *cru*: *A. cruentus*, *hyp*: *A. hypochondriacus*

2.3 Discussion

Crop populations that we observe today are the result of different evolutionary processes. While selection and demographic changes have been extensively studied, gene flow and its role in the fitness of crops has only received attention in recent years. This is partially due to technical advances in plant genomics, but might also have resulted from the conceptual assumption of a linear process from one wild ancestor. The complex makeup of modern grain amaranth shows that gene flow between crop populations, even over long geographic distances, was prevalent (Figure 2). Gene flow between crop lineages of species that were domesticated multiple times has also contributed to diversity in rice (Yang et al., 2012), tomato (Razifard et al., 2020) and common bean (Rendón-Anaya et al., 2017). Not only does such gene flow between closely related crop populations occur, it is also heterogeneous between individuals and along the genome (Figure 3).

A potential reason for lower genetic exchange between specific pairs of grain amaranth could have been the reported difference in chromosome number between *A. cruentus* with 17 chromosomes in comparison to 16 chromosomes in the other 4 species (EJ and Poggio, 1994; Ma et al., 2021). Yet, this would only lead to infertile F_1 plants, rather than necrotic, non-viable plants that die in the seedling stage. Instead, *A. cruentus* formed fertile hybrids with two out of three *A. hypochondriacus* accessions, suggesting incompatibility not because of difference in chromosome numbers but rather a genetic incompatibility (Figure 6). Our crossing experiment revealed differential genetic incompatibility between grain amaranth species, consistent with previous observations on interspecific hybrid necrosis (Gupta and Gudu, 1991). A potential one-locus underdominance model of hybrid incompatibility would require strong genetic drift in both populations (Wu and Ting, 2004), which could be the result of previously demonstrated domestication bottlenecks in grain amaranth (Stetter et al., 2020). The observed reproductive barrier could also be the result of a Dobzhansky-Muller incompatibility (Muller, 1942), including more than one locus. Given the large geographic distance between incompatible crop species (*A. caudatus* in South America and *A. cruentus* in Mesoamerica) the reproductive barriers likely evolved through neutral processes rather than selection against gene flow. The incomplete barrier between *A. cruentus* and *A. hypochondriacus* might allow further insights into the progression of reproductive isolation during crop domestication (Tenailon et al., 2023). The genetic mechanism for incompatibility warrants further investigation, as this has practical implications for potential hybrid breeding using different crop species as heterotic pools. The complete compatibility between *A. hypochondriacus* and *A. caudatus* is reflected in higher gene flow signals than between these species and *A. cruentus* (Figure 2 and 6). Therefore, *A. hypochondriacus* and *A. caudatus* would be the most promising heterotic pools for future amaranth breeding.

Given the high prevalence of incompatibility between *A. cruentus* and *A. caudatus*, gene flow might

have occurred early during the 8,000 year-long domestication history if strong reproductive barriers only developed later in the process. If the hybrid incompatibility arose at the advent of amaranth domestication, it can be expected that gene flow between *A. cruentus* and *A. caudatus* was likely beneficial, given the high fitness disadvantage of F_1 hybrids (Janzen et al., 2019; Aguillon et al., 2022). We found a low but significant number of introgressed selective sweeps that might represent such beneficial gene flow between amaranth crop species. Currently, there are only a few domestication-related QTL known in grain amaranth that would allow linking introgressed regions to phenotypic changes during domestication. The previously reported QTL for the seed color change during amaranth domestication did not show signals of introgression consistent with the previously reported repeated selection for the trait in the three crop species (Stetter et al., 2020). More quantitative genetic and functional analyses could reveal additional QTLs that could indicate the adaptive potential of introgressed regions between grain amaranths.

Adaptive gene flow from wild relatives into crops has previously been associated with environmental adaptation in crops. For instance, in maize, the introgression of the wild relative *Zea mays spp. mexicana* has been associated with the adaptation of maize to highland conditions and colder climates (Wang et al., 2017). Recent work even suggests a prevalent role of *Zea mays spp. mexicana* in the domestication of maize (Yang et al., 2023). While we cannot associate gene flow from wild relatives with positive selection, we found decreased genetic load in regions that were introgressed from wild relatives (Figure 5). This could be due to higher effective population size and higher genetic diversity of wild relatives (Stetter et al., 2020). The wild ancestor *A. hybridus* showed lower genetic load than the domesticates (Figure 5), which might be the result of less demographic change during the recent past (Stetter et al., 2020). Post-domestication gene flow between crops and their wild ancestor could consequently reduce the frequency of deleterious alleles. A similar correlation of genetic load and gene flow from a wild relative has also been shown in maize and sunflower, where gene flow regions from the wild relative showed reduced genetic load (Wang et al., 2017; Huang et al., 2023). Similarly, work in humans has shown that gene flow from a relative with a small population size (Neanderthal) into a population with a larger population size (modern humans) led to increased genetic load (Harris and Nielsen, 2016), as we observe for gene flow from domesticated amaranths into *A. hybridus*. The accumulation of genetic load in populations with small effective population sizes can even lead to the extinction of populations or species as a whole (Rogers and Slatkin, 2017). Hence, gene flow from relatives with large population size not only provides adaptive variation but can also lead to the evolutionary rescue of the small population (Carlson et al., 2014). Despite the amelioration of genetic load through gene flow with wild relatives, crop-wild hybrids are expected to perform poorly as crops. Gene flow, therefore, needs to have an overall beneficial effect that is higher than the destruction of domestication traits (Stetter, 2020). This might be particularly possible in crops like grain amaranth where the domestication syndrome is only weakly pronounced (Stetter et al.,

2020).

For crops that maintain high gene flow with their relatives and are phenotypically less differentiated from wild plants, as is the case for grain amaranth, the borders between wild and domesticate might be fluid. Strong gene flow between wild and domesticated crops might also allow the domesticate to return to the wild and be viable without human intervention again. Such feralization has been observed for a number of plant and animal species (Gering et al., 2019). We found particularly strong signals of gene flow between grain amaranth and its wild relatives in South America (Figure 2B). The close relationship between wild species and crop in South America has led to different hypotheses for the domestication of *A. caudatus* suggesting *A. quitensis* as potential wild ancestor (Sauer, 1967b). While this cannot be completely ruled out, previous work using genome-wide makers data suggested *A. hybridus* as ancestor for all three grain amaranths (Kietlinski et al., 2014; Stetter et al., 2020). The high and genome-wide equally distributed signal of gene flow between *A. caudatus* and *A. quitensis* (Figures 2 and 3) together with the low population size and signs of a strong population bottleneck in *A. quitensis* (Stetter et al., 2020) might indicate a feralized status of this species. The clarification of the status of *A. quitensis* will need further work with multiple populations of this species, local crop and wild relatives.

Overall, we show that the relationships between species are beyond linear, with exchanges between populations despite large geographic distances and the reintroduction of genetic material that was potentially lost during speciation. Even with observed genetic incompatibilities and high genetic differentiation between the crop species, we found strong signals of gene flow between grain amaranths and between the crops and their wild relatives. Recurrent gene flow from the wild relative into the crops might have allowed evolutionary rescue, counteracting the loss of diversity, but likely hindered the fixation of domestication traits leading to the incomplete domestication syndrome observed today for grain amaranth.

2.4 Materials and Methods

We studied whole genome resequencing data of 108 domesticated and wild amaranth accessions. The raw reads are available from European Nucleotide Archive (project numbers PRJEB30531) (Stetter et al., 2020). The accessions included the three domesticated amaranth; 33 *A. caudatus* L. (*caudatus*), 21 *A. cruentus* L. (*cruentus*) and 21 *A. hypochondriacus* L. (*hypochondriacus*); as well as 5 wild *A. hybridus* L. from Central America (*hybridus_CA*), 4 *A. hybridus* L. from South America (*hybridus_SA*), and 24 *A. quitensis* Kunth. (*quitensis*) (Table S5). *A. tuberculatus* was used as outgroup for the study (ERR3220318), (Kreiner et al., 2019). Raw reads were aligned to amaranth reference genome (Lightfoot et al., 2017b) using bwa-mem2 (v 2.2.1) (Vasimuddin et al., 2019).

Variant calling

For variant calling, we utilized ANGSD (v.0921) (Korneliussen et al., 2014b), with `-ref A. hypochondriacus V2.1` reference genome (Lightfoot et al., 2017b) - `doCounts 1`, `doGeno 3` `dovcf 1`, `gl 2`, `dopost 2`, `domajorminor 1` and `anddomaf 1`. We filtered for missing data and mapping quality using `minInd 73` (max 30 missing data), `minQ 20`, `minMapQ 30`, `only_proper_pairs 1`, `trim 0`, `SNP_pval 1e-6`, `setMaxDepthInd 150` and `setminDepth 73`. The resulting VCF file was phased using Beagle (v 5.2) (Browning et al., 2021) using default parameters. For linkage disequilibrium (LD) pruning, we used Plink (v 1.9) (Purcell et al., 2007) using windows of 50kb with 5kb steps and a r^2 threshold of 0.3. The resulting VCF file had a total of 13,330,082 sites.

Gene flow analysis

We inferred gene flow between populations using D-statistic implemented in ANGSD (v.0921) (Korneliussen et al., 2014b). We inferred population-wide statistics with the `abbababba` function for calculations of D per individual and `abbababba2` for calculations between populations. For both tests, *A. tuberculatus* was used as outgroup (H4). Only trees with a significant Z-score (absolute value above 3) were included in the results. For fine-scale analysis along the genome, we employed Dsuite (Malinsky et al., 2021-02). We used the function `Dinvestigate` in windows of 100 SNPs to calculate D between trios along the genome. We also utilized the function `Dtrios` to verify the concordance of the global genome with the results obtained from ANGSD. To overlap regions with gene flow between crop species with selection signals, windows with significant signals of gene flow (1% outlier values) were overlapped with selective sweeps signals in the recipient population. Selective sweeps were previously identified in Gonçalves-Dias and Stetter (2021). The overlaps were tested for significance using a hypergeometric test (`pyhper` function in R 4.2).

Topology inference

We used Twisst (Martin and Van Belleghem, 2017) to infer the topology of each trio along the genome in windows of 100 SNPs, utilizing *A. tuberculatus* as an outgroup. For each window, a topology is assigned and a summary of the proportional windows in which each topology appeared is then obtained. This inference allows a blind observation of the relationship between species. In the case, where topologies that differ from a neutral expectation are present in high proportions suggests gene flow between species. We inferred the topology for trios, between which a putative gene flow signal was identified.

Local Ancestry inference

We inferred local ancestry for each individual using `finestructure v4.1` (Lawson et al., 2012). We used a uniform recombination rate generated with the perl script `makeuniformrecfile.pl`. The program was run using parameter `-f 0 0` which considers the populations and iterates through all individuals. For each individual, all five species (including individuals of the same species) were used as donors, which allowed to differentiate ancestry from all the species. We further assigned the most likely donor for each genomic region of an individual. The donor for each position per individual was assigned based on the most likely donor population with a likelihood larger than 0.5. Sites that could not meet these likelihood thresholds were called "ambiguous". Using these thresholds, the proportion of donated region per individual was calculated. In addition, the proportion along the genome within each population was summarized.

2.4.1 Demographic modeling

To estimate whether the identified scenario of gene flow fits best to our data we used simulations using `Fastsimcoal2` (Excoffier et al., 2013). The joint site frequency spectrum (SFS) was generated using non-coding SNPs having no missing value in any of the individuals and a minimum coverage of five reads using a python program `easySFS` (<https://github.com/isaacovercast/easySFS>). First, the program was run on preview mode (`-preview`) to identify the true sample size and the best sample size selected was used for the projection (`-proj`) to generate the joint SFS. Three different models namely, two-population split, two-population split with one migration event and two-population split with continuous gene flow (Figure S2) were applied to the observed joint SFS. The models were compared using Akaike's Information Criterion (AIC). The best parameter estimate was calculated based on 100 independent runs with 200,000 coalescent simulations and 40 cycles of likelihood maximization algorithm. The 95 percent confidence interval of each parameter was estimated based on 50 non-parametric bootstrapping datasets. Each of the 50 bootstrapped datasets was run 50 times to estimate the best run. These 50 best run parameters were then used to estimate the confidence interval using the `boot` package in R (Canty and Ripley, 2017).

Genetic Load

We used Genomic Evolutionary Rate Profiling (GERP) (Davydov et al., 2010) scores to account for the effect of deleterious alleles at each site. We aligned 15 repeat-masked genomes of angiosperm species spanning a large taxonomic range to the reference genome of *A. hypochondriacus* (v2.1) (Lightfoot et al., 2017b). We followed the pipeline of Wu et al. (2022). Briefly, we aligned genomes of 16 divergent

species: *Beta vulgaris* EL10 1.0, *Brachypodium distachyon* v2.1, *Chenopodium quinoa* v1.0, *Glycine max* Wm82.a4.v1, *Helianthus annuus* r1.2, *Medicago truncatula* Mt4.0v1, *Mimulus guttatus* v2.0, *Oryza sativa* v7.0, *Phaseolus vulgaris* v2.1, *Populus trichocarpa* v4.1, *Setaria viridis* v2.1, *Solanum lycopersicum* ITAG4.0, *Sorghum bicolor* v3.1.1, *Spinacia oleracea* (Monoe Viroflay) and *Vitis vinifera* v2.1 from phytozome (Goodstein et al., 2012) to the *A. hypochondriacus* reference genome using the LAST aligner (Kiełbasa et al., 2011). The tree topology for the species was extracted from the NCBI-phylogeny using the ete3 toolkit (v3.1.2) (Huerta-Cepas et al., 2016). Phylofit from phast package (Siepel et al., 2005) was used to calculate the branch length of the tree along with four-fold degenerate sites from *A. hypochondriacus* reference annotation file to generate a neutral model. All pairwise alignments were merged using ROAST (<https://github.com/multiz/multiz>) (Blanchette et al., 2004). GERP++ (Davydov et al., 2010) was used to calculate the GERP scores for each site using `gerpcol` with `-j` option that projects out the reference genome to avoid bias in the calculation. All sites with negative GERP values were set to 0 as negative values are not informative and misleading. The ancestral allele for each site was defined on the basis of three outgroup species closest to *Amaranthus* in the phylogenetic tree, i.e., *Beta vulgaris*, *Chenopodium quinoa* and *Spinacia oleracea*. Variants among these three species were called from the roast multiple alignment file (maf) using maffilter (version 1.3.1) (Dutheil et al., 2014). Sites not covered by any of the three outgroup species were removed. For the remaining sites, the major allele among the three species was called the ancestral allele. In case of discrepancy for a majority rule, the allele for *Spinacia oleracea* was called as the ancestral allele.

From a total of 13.2 million SNPs, we extracted 1,429,744 sites for which the GERP score and the ancestral alleles could be assigned. We then polarised the SNPs and removed sites where neither the reference nor the alternate allele matched the ancestral one. This yielded a total of 1,145,566 SNPs that were used for genetic load calculation. The GERP score from sites having derived alleles was then summed to calculate the total genetic load using an additive model. The genetic load was calculated for each accession individually. The significance of variation for the load in each population was then analyzed using an ANOVA followed by TukeyHSD in R (<https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/TukeyHSD>).

To account for the reference bias, the GERP score was also calculated using *A. cruentus* genome (Ma et al., 2021) as reference, using the same method as described above. The variants file of individuals that used *A. hypochondriacus* as reference was lifted to the *A. cruentus* genome using liftover module of maffilter (version 1.3.1) (Dutheil et al., 2014). The individual genetic load was then calculated using GERP score from *A. cruentus* genome for each lifted site as described above.

To overlap genetic load and selective sweeps, the sweep regions for each of the populations were overlapped with the sites with GERP score. Variants from those regions were extracted for the individuals

of the respective populations. Any other regions of the genome that were not under the sweep region were considered as control region. To account for the difference in the sizes of the sweep and control regions, we divided the total load by the total number of sites used in the analysis. Differences in genetic load between sweep and non-sweep regions were then tested using a t-test for each population.

In order to calculate the introgressed load, GERP scores were summed for sites having the derived allele donated by a different population. The introgressed load was expressed as a ratio of load contributed by the introgressed species to the percentage of introgression sites. A value greater than one predicts the contribution of a higher load due to gene flow. The significance of deviation from the expected value was analyzed using one-sample t-test against the null-expectation of equal contribution of 1.

Experimental hybridization between grain amaranth species

We selected genetically and morphologically defined accessions for each of the three crop species (*A. caudatus*, *A. cruentus*, and *A. hypochondriacus*) to assess their cross-compatibility. We crossed 9 parental accessions (three per species) to create 25 inter- and intraspecific combinations and examined multiple crosses per combination (Table S3). The parental lines were previously selfed for at least three generations to ensure homozygosity. As the three grain amaranths are mostly selfing, we hand-emasculated the female parent and bagged parents together. Successful crossing was ensured by PCR using diverging primer pairs (Table S4). To determine the hybrid survival rate, the hybrids were grown alongside their parental accessions in a greenhouse in Cologne (Germany) under long day conditions (16h light, 8h dark) at 25°C. We evaluated the survival of hybrid plants from at least three offsprings per cross. We considered a cross as "lethal" when all F₁ seedlings died within 20 days after planting.

2.5 Data Availability

Genomic data is available through Stetter et al. (2020) and the associated ENA project. All scripts used in the analysis are available on <https://github.com/cropevolution/GeneFlowLoadRescue>.

2.6 Author contribution

MGS conceived the study. JGD processed the data and conducted genome-wide analysis of gene flow and local ancestry inferences. AS performed genetic load analysis and demographic modeling. CG performed experimental crosses and executed the incompatibility study together with JGD. JGD, AS and CG prepared figures and tables. MGS, JGD and AS wrote the manuscript. All authors discussed the results, edited and approved the manuscript.

2.7 Competing interests

The authors declare that they have no competing interests.

2.8 Acknowledgments

We acknowledge funding by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC-2048/1 – Project ID 390686111 and grant STE 2654/5 to MGS by the DFG.

3 PopAmaranth: a population genetic genome browser for grain amaranths and their wild relatives

Abstract

The combination of genomic, physiological, and population genetic research has accelerated the understanding and improvement of numerous crops. For non-model crops the lack of interdisciplinary research hinders their improvement. Grain amaranth is an ancient nutritious pseudocereal that has been domesticated three times in different regions of the Americas. We present and employ PopAmaranth, a population genetic genome browser, which provides an accessible representation of the genetic variation of the three grain amaranth species (*A. hypochondriacus*, *A. cruentus*, and *A. caudatus*) and two wild relatives (*A. hybridus* and *A. quitensis*) along the *A. hypochondriacus* reference sequence. We performed population-scale diversity and selection analysis from whole-genome sequencing data of 88 curated genetically and taxonomically unambiguously classified accessions. We employ the platform to show that genetic diversity in the water stress-related MIF1 gene declined during amaranth domestication and provide evidence for convergent saponin reduction between amaranth and quinoa. PopAmaranth is available through amaranthGDB at amaranthgdb.org/popamaranth.html

3.1 Introduction

Genome sequencing, genome-assisted breeding, and molecular breeding techniques have accelerated the improvement of numerous major crops (Wallace et al., 2018; Lemmon et al., 2018). The availability of genome-wide diversity data of crops and their wild relatives has allowed to identify and study candidate genes of agronomic significance (Hufford et al., 2012; Huang et al., 2012; Wang et al., 2020a). These candidates can then be validated through molecular genetics (Ross-Ibarra et al., 2007; Fernie and Yan, 2019; Sedeek et al., 2019; Wang et al., 2020a). To facilitate the interdisciplinary use of population genetic results, it is essential to provide summary statistics in an intuitive and user-friendly way.

Different platforms have been developed to make genomic resources available across disciplines and have enabled the integration of complementary research areas (Lawrence et al., 2004; Alonso-Blanco et al., 2016; Jin et al., 2013). Online genome browser platforms such as Ensemble (Bolser et al., 2016) and Phytozome (Goodstein et al., 2012) have become a standard interface to interact with genome sequences and annotations and are used across research fields. Genome browsers provide access to reference genome sequences and gene annotations for numerous plant species but most browsers only provide data for a

single reference individual per species. Species-specific browsers include sequence data and variant calls for a large number of individuals (e.g., Lawrence et al., 2004; Dash et al., 2016; Krishnakumar et al., 2015; Mansueto et al., 2017; Kudo et al., 2017), but do not allow a direct inference of a population scale genome-wide diversity across related species. Population genetic genome browsers that provide population-scale summary statistics are available for only a few non-plant model species (Casillas et al., 2018). For plant and crop species, in particular minor crops, such resources are currently unavailable.

Novel and under-utilized crops have a high potential to contribute to sustainable food production, as many such crops are tolerant to abiotic and biotic factors and are of high nutritional value (Mayes et al., 2012). Amaranth is an under-utilized crop that has been cultivated for its grains as pseudocereal and its edible leaves as a vegetable (Sauer, 1967a; Joshi et al., 2018). Three grain amaranth species, *Amaranthus caudatus*, *A. cruentus* L., and *A. hypochondriacus* L., have been domesticated for their grain from a common wild ancestor, *A. hybridus* L. (Stetter et al., 2020). Another wild relative, *A. quitensis* Kunth, is suspected to be involved in the domestication of the South American *A. caudatus*, although its role and contribution to the crop remain unclear (Stetter et al., 2017a, 2020). The repeated domestication of amaranth presents an interesting model to study genetic parallelisms along selection gradients, and the combination of genomics, quantitative genetics, and molecular dissection of gene function has a high potential to improve grain amaranth.

First resources that allow the functional study of traits have been developed for amaranth. On the one hand, numerous genomic resources, including a high-quality reference genome (Lightfoot et al., 2017a) and a transcriptome (Clouse et al., 2016), genome-wide marker data (Mallory et al., 2008; Stetter et al., 2017a, 2020) and QTL regions for different traits (Lightfoot et al., 2017a; Stetter et al., 2020) have been identified. On the other hand, a number of molecular methods have been adapted for the crop, including molecular gene function identification (Massange-Sanchez et al., 2016), state-of-the-art transient 'hairy' roots expression systems (Castellanos-Arévalo et al., 2020), and stress physiology assays (Parra-Cota et al., 2014; Massange-Sanchez et al., 2015). Combined, these resources can elevate amaranth research and improvement if results and data are available and accessible for researchers across disciplines.

Here, we present PopAmaranth, an interactive genome-wide population genetic browser for amaranth. PopAmaranth facilitates browsing a number of population genetic summary statistics and selection signals, gene annotation, and variant calls of the three grain amaranths and two wild relatives along the amaranth genome. We defined a curated set of 88 morphologically and genetically identified samples with whole-genome sequencing data to represent the five populations. Currently, PopAmaranth provides three categories of summary statistics, namely genetic diversity, population differentiation, and selection signals, plus variant calls and annotation tracks, in a total of more than 40 tracks. We show how the tool allows a user-friendly way to screen evolutionary signals for candidate genes and compare

them between populations by identifying selection signals in a stress gene previously identified in one of the grain amaranths and in an ortholog quinoa domestication gene that shows convergent signals of selection in amaranth. PopAmaranth is embedded in amaranthGDB and is accessible from amaranthgdb.org/popamaranth.html.

3.2 Methods

3.2.1 Data and filtering

We used whole-genome sequencing data of 116 accession from five amaranth species, including the three grain amaranths (24 *A. hypochondriacus*, 24 *A. cruentus*, and 34 *A. caudatus* samples) and their two wild relatives, 9 *A. hybridus* and 25 *A. quitensis* (Stetter et al., 2020, Table S1). The sequencing reads were aligned to the *A. hypochondriacus* reference sequence V 2.0 (Lightfoot et al., 2017a).

We performed principal component analysis (PCA) on the full set of accessions to remove individuals with ambiguous species clustering using PCAngsd (version 0.982) (Meisner and Albrechtsen, 2018) followed by prcomp to calculate principal components. We manually excluded samples that did not genetically cluster with the morphologically designated species information in their passport data after visual evaluation of the first three PCs (Figure S1). We only used bam files of remaining individuals for summary statistic estimation and subset the VCF file from Stetter et al. (2020) using VCFtools 0.1.16 (Danecek et al., 2011) to only include sites that segregate in this set.

We calculated the site allele frequency likelihood based on individual genotype likelihoods for each of the five species using the `-doSaf 1` function on ANGSD (version 0.930) (Korneliussen et al., 2014a). We removed sites with a minimum map quality below 30, minimum base qscore below 20, and a flagstat (Li et al., 2009) above 255, keeping only primary reads (`-doSaf 1, -GL 2, -remove_bads 1, -minMapQ 30, -minQ 20`). In addition, we removed all sites with more than 66% missing values (`-minInd=1/3*n`).

3.2.2 Population genetic browser tracks

Using `realSFS saf2theta` functions on ANGSD, we calculated the folded site frequency spectrum and estimated per site thetas (population scaled mutation rate). Consequently, we calculated nucleotide diversity (π) and Wu and Watterson estimator (θ) in non-overlapping windows of 5000 bp using the `do_stats` function of ANGSD. We only kept windows with more than 30% of the sites called in a given window.

To data as browser tracks we converted the files to bigWig format using UCSC `bedgraphtobigwig` (Kent et al., 2010). Within the genome browser, a yellow horizontal line denotes the genome-wide mean

for each of the summary statistics, values below the mean are shown in red and above the mean in blue and we indicated the strength of deviation by adding dark gray and light grey shadings for one and two standard deviations from the mean, respectively.

We calculated pairwise Weir-Cockerham F_{st} (Wright, 1950) as a measure for genetic differentiation for each pair of populations using ANGSD (Korneliussen et al., 2014a). We used these values as input to calculate pairwise F_{st} in non-overlapping windows of 5000bp along the genome.

We employed ANGSD (Korneliussen et al., 2014a) with the parameters described above for π and θ to calculate Tajima's D in non-overlapping 5 kb windows. Using the nucleotide diversity estimated for each of the species, we calculated relative nucleotide diversity. We divided π for each of the domesticated species (*A. caudatus*, *A. cruentus*, and *A. hypochondriacus*) by π of their wild ancestor, *A. hybridus*. We only used windows where both species had data after filtering for the number of genotyped sites.

Variant based statistics were calculated based on the sub-sampled VCF data from Stetter et al. (2020). We used the scikit-allel python library (<https://doi.org/10.5281/zenodo.597309>) to calculate per site heterozygosity statistics (H_{exp} , H_{obs} , and F) for each of the five populations. We applied Raised Accuracy in Sweep Detection (RAiSD) (Alachiotis and Pavlidis, 2018) with default setting (20 SNP windows) on the subset VCF data from Stetter et al. (2020) to detect signals of selective sweeps within each population. We considered windows on the top 1 % μ values as outliers and under positive selection (*A. caudatus*: 17650 windows; *A. cruentus* 16546; *A. hypochondriacus*: 17932 *A. hybridus*: 43415; and *A. quitensis*: 15854). We merged all overlapping windows to create stretches of selective sweeps.

3.2.3 Browser implementation and annotation

We provided access to the summary statistics described above as an interactive tool through JBrowse 1.16.9 (Skinner et al., 2009). We added the reference sequence and gene annotation, including exons, introns, CDS, mRNA, and UTRs from Lightfoot et al. (2017a) available through Phytozome (Goodstein et al., 2012). For each summary statistic a color gradient summary plot combining all species was added. Further, we added the "Variant" category, providing variant data for biallelic SNPs within each species from Stetter et al. (2020) (not including variants fixed between populations).

3.2.4 PopAmaranth application to candidate genes

We downloaded the sequence of the water stress-related MIF1 gene reported in Huerta-Ocampo et al. (2011) from the NCBI database and used BLASTn (Altschul et al., 1990) to identify the gene ID in the *A. hypochondriacus* V2 reference sequence on Phytozome. Using the same procedure, we studied the triterpene saponin biosynthesis activating regulator-1 (TSAR-1) gene from *Chenopodium quinoa* (Jarvis

et al., 2017).

3.3 Results

3.3.1 Sample filtering

Amaranthus species are difficult to taxonomically classify because of their high morphological similarity (Sauer, 1967a). Therefore, we sub-sampled the original dataset from Stetter et al. (2020) based on the genetic clustering in the PCA and species delimitation in Germplasm Resources Information Network (GRIN). We selected each species according to their clustering in the first three principal components (Figure S1). After filtering, our sample consisted of 88 genetically and morphologically defined samples representing the five species, with 28 individuals classified as *A. caudatus* L., 21 *A. cruentus* L., 18 *A. hypochondriacus* L., 12 *Amaranth quitensis* Kunth, and 9 *A. hybridus* L. (Figure 7 and table S1).

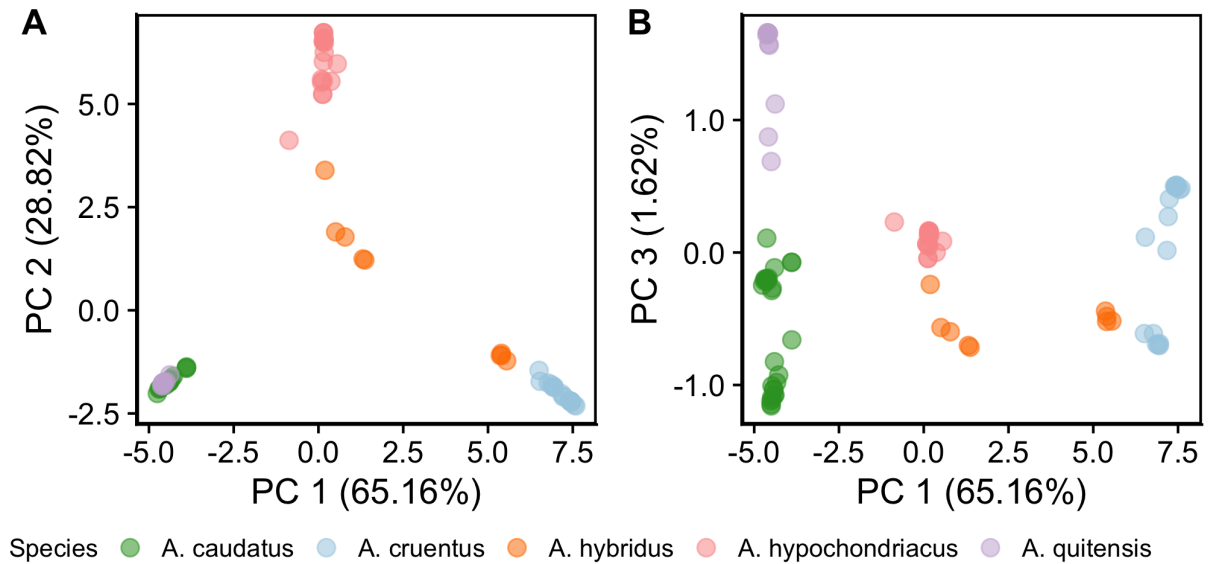


Figure 7: **Principal Component Analysis with filtered samples.** Each dot represents each of the 88 samples. *A. caudatus* (green), *A. cruentus* (blue), *A. hybridus* (orange), *A. hypochondriacus* (rose), *A. quitensis* (purple). Axis show the percentage of variance explained by each principal component

3.3.2 Categories and Tracks

We created PopAmaranth relative to the high-quality *A. hypochondriacus* reference genome (Lightfoot et al., 2017a) and added the gene annotation as functional guide. We calculated nine summary statistics from whole-genome sequencing data for each of the five species. The tracks are grouped into five categories, namely annotation, differentiation, diversity, selection, and variant calls (Table 1 and S2). Each

category includes tracks one color gradient summary track combining data of a summary statistic for all species.

Table 1: Tracks available in PopAmaranth

Track	Description
Annotation	
Reference Genome v2.0	<i>Amaranthus hypochondriacus</i> reference genome v2.0 (Lightfoot et al., 2017a)
Gene Annotation v2.1	<i>Amaranthus hypochondriacus</i> gene annotation with subfeatures, including CDS, mRNA and UTRs
Differentiation	
F_{st}	Fixation Index, average pairwise differences Weir and Cockerham (1984)
Diversity	
Wu & Watterson θ	Estimator of genetic diversity in a population (Watterson, 1975)
Expected heterozygosity	Expected heterozygosity for a SNP under Hardy-Weinberg equilibrium
Observed heterozygosity	Observed heterozygosity for a SNP genotype.
Inbreeding coefficient	Inbreeding coefficient (F) for each variant
Nucleotide diversity (π)	Nei's nucleotide diversity (Nei and Li, 1979)
Selection	
Tajima's D	Scaled difference between the mean number of pairwise differences and the number of segregating sites Tajima (1989)
Relative nucleotide diversity	Ratio of nucleotide diversity between a domesticated species and their wild ancestor (<i>A. hybridus</i>)
Selective Sweep (RAiSD (μ))	μ statistic for selective sweep detection
Variant Call	
VCF	Called SNPs with a given species and their genotype frequency

Differentiation

Tracks in the differentiation category represent all pairwise F_{st} comparisons in 5 kb windows. The genome-wide pairwise F_{st} ranged from 0.17 between *A. caudatus* and *A. quitensis* to 0.68 between *A. caudatus* and *A. cruentus*. As observed before, F_{st} between crop species was higher than between the crops and their wild ancestor for *A. caudatus* and *A. hypochondriacus* (Stetter et al., 2020). Although, we found higher F_{st} between *A. cruentus* and *A. hybridus* (0.69) than between *A. cruentus* and *A. hypochondriacus* (0.57).

Diversity

Genetic diversity patterns along the genome can give insights into the evolutionary history of a population. Hence, we calculated several diversity statistics along the genome. Inbreeding coefficients and expected and observed heterozygosity are reported on a per-site basis for each SNP that segregated within a population. In addition to SNP-based statistics, we provide windowed diversity measures, including Wu & Watterson θ and nucleotide diversity π in 5 kb non-overlapping windows. Consistent with previous findings, the three grain amaranths had a lower mean π (0.005-0.010) compared to their wild ancestor *A. hybridus* (0.019) (Stetter et al., 2020). Wu & Watterson θ was also lower for domesticated amaranth species (0.004-0.007) compared to *A. hybridus* (0.023).

Selection

We calculated three different summary statistics to detect signals of selection along the genome. Tracks displaying Tajima's D were calculated in 5 kb windows for each species. Tajima's D was higher for domesticated species (1.443 in *A. caudatus*, 1.773 in *A. cruentus*, and -0.105 in *A. hypochondriacus*) than for their wild ancestor *A. hybridus* (-0.597), indicating a domestication bottleneck. *A. quitensis* had a mean Tajima's D of 2.037 also suggesting a recent population contraction.

We employed RAI_{SD} to detect signals of selective sweeps in 20 SNP windows within each species. The top 1% of all windows were considered outliers and suggest regions of positive selection. After merging adjacent outliers, we found 973 non-overlapping windows with positive selection signals in *A. caudatus*, 1,096 in *A. cruentus*, 1,121 *A. hypochondriacus*, 2,452 *A. hybridus*, and 1,275 windows in *A. quitensis*. To investigate the signal of domestication-related selection, we added the relative nucleotide diversity between each crop and their wild ancestor *A. hybridus* in 5 kb windows. While the genome-wide π was lower for all three crops (see "Diversity"), relative π allows to visualize deviations from this genome-wide mean and detect outlier signals in individual regions.

Variant Calls

Individual variants give access to an individuals' genotype. Molecular biologists might be interested in evaluating natural alleles of a gene of interest, and plant breeders could use individuals with specific variants to enrich their gene pools. We provide variant data for all five species representing their genotype frequency within the population. Each variant track only displays variants within the given population (not including fixed variants between populations). A total of 4,961,210 variants for *A. caudatus*, 4,075,368 for *A. cruentus*, 4,551,278 for *A. hypochondriacus*, 12,238,589 for *A. hybridus*, and 2,342,505 for *A.*

quitensis along the genome are available.

3.3.3 PopAmaranth case study

To show the utility of PopAmaranth, we evaluated the evolutionary signals for a gene that was molecularly shown to be involved in the response of *A. hypochondriacus* to water stress (Huerta-Ocampo et al., 2011). We found that MIF1 (AH-017582) showed lower nucleotide diversity, decreased expected heterozygosity, and a relative nucleotide diversity below the genome-wide average in all three grain amaranth species. Also, we identified a selective sweep in *A. hypochondriacus* around this gene (Figure 8). Our findings combined with the "Variant call" tracks in the browser allow to select accessions with contrasting genotypes to identify the causal allele for the expression difference in *A. hypochondriacus* and compare wild and domesticated amaranth for their drought response.

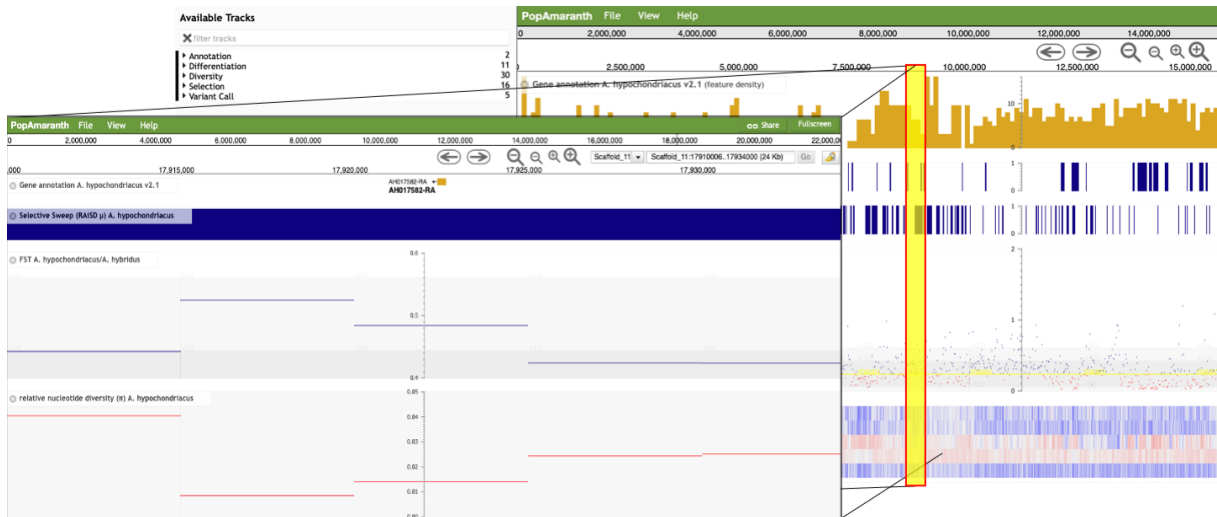


Figure 8: **PopAmaranth screen view.** Background panel: Zoomed out user view along a chromosome. Search field provides access to genome positions or gene names. Front panel: example is illustrated with a zoom-in region for the water-stress related MIF1 gene (AH-017582). The blue bar shows a selective sweep in *A. hypochondriacus*, which is supported by high F_{st} between wild ancestor and crop and low π in the crop

In addition to the amaranth specific use, PopAmaranth facilitates the evaluation of hypothesis beyond the species. To show its utility to study convergent selection signals across distant families, we evaluated population genetic signals around the amaranth ortholog to the triterpene saponin biosynthesis activating regulator 1 - TSAR1 (AH-019562), a key regulator for seed saponin content in *Chenopodium quinoa* (Jarvis et al., 2017). We found signals of selective sweeps in the three grain amaranth species. Furthermore, the relative diversity compared to the wild ancestor was below the genome-wide mean, suggesting selection during amaranth domestication (Figure S2).

3.4 Discussion

Over the last decades, large-scale population genomic data revealed insights into the evolution and adaptation of crops. Providing access to results in a user-friendly and interactive way opens paths to better integrate data from different research areas. Our population genomic genome browser, PopAmaranth, aims to provide such an intuitive tool for amaranth population genetic results. The inclusion of five different species involved in the crop domestication history of facilitates hypothesis testing along this evolutionary gradient.

For other plant species, i.e., maize (Lawrence et al., 2004), tomato (Fernandez-Pozo et al., 2015), and arabidopsis (Alonso-Blanco et al., 2016) accessible platforms of genomic and evolutionary data are integral parts of the research communities. We hope that PopAmaranth and the higher-level framework amaranthGDB will help establish an amaranth community that benefits from the interdisciplinary exchange.

Our results show how PopAmaranth can be employed to add an evolutionary perspective to different molecular questions. We identify previously unknown signals of selection in stress-related MIF1 gene, which might have been under selection during amaranth domestication. In most crops, domestication led to a reduction in stress resilience compared to their wild ancestors. Hence, the reduction in diversity might represent selection against the tolerant allele to free resources for increased crop productivity (Koziol et al., 2012). Our browser allows the selection of genotypes with different alleles within grain amaranths and in wild amaranth, enabling the identification of stress-tolerance alleles and potentially the reintroduction of such alleles into breeding programs.

On a broader scale, PopAmaranth also facilitates the comparison of convergent adaptation signals between more distant taxa. For instance, our finding of convergent selection between quinoa and amaranth in a saponin-related gene suggests that in both quinoa and amaranth the saponin content was reduced to improve the palatability of the grains (Jarvis et al., 2017). Saponins confer toxicity to protect wild plants against birds but reduce the nutritional quality of seeds for human consumption and animal feed (Oleszek et al., 1999; Mroczek, 2015). Hence, our platform allowed to identify the convergent selection between the two pseudocereals, demonstrating its utility to evaluate selection signals across taxa. This is of particular use for close relatives of weedy *Amaranthus* species that are of evolutionary and agronomic interest and have been aligned to the same reference genome used in our browser (Montgomery et al., 2020).

We aimed for a generalized usage of diversity and differentiation estimates. Therefore, we only selected unambiguous samples of each species, based on morphological and genetic classifications. A clear grouping is crucial for a reference tool, as misclassified samples would confound population-wide

signals (Rieseberg and Wendel, 2004). Our sub-sampling approach is conservative regarding genetic diversity, as it excludes more differentiated individuals from the analysis. Reported values of genetic differentiation (F_{st}) between species could be inflated due to the lack of intermediate individuals. The increased differentiation by sub-sampling potentially led to the higher F_{st} value between *A. cruentus* and *A. hybridus* compared to previous results (Stetter et al., 2020). While there is a trade-off between including additional individuals and the potential for undiscovered diversity, our goal was a defined and distinguished set of samples representing each species. The inclusion of only core individuals of each species further allows the comparison and classification of less distinct individuals using our set.

Altogether, we incorporated a well-defined set of individuals with congruent data filtering to estimate population-wide diversity statistics for the three grain amaranth species and two wild relatives. The identification of selection signals in candidate genes within amaranth and beyond shows the utility of the browser for a range of researchers. PopAmaranth and the amaranthGDB platform will help build and grow the amaranth research community and facilitate interdisciplinary research to ultimately improve the crop.

3.5 Availability

PopAmaranth is available at <https://amaranthgdb.org/popamaranth.html>. A static version of the browser and data in table format can be found at: <https://doi.org/10.6084/m9.figshare.13340798.v1>. Code is available at <https://github.com/cropevolution/PopAmaranth>. Tracks data for the region in observation can be downloaded directly from the tracks options.

3.6 Acknowledgments

We thank the RRZ at University of Cologne, for hosting PopAmaranth, Benedict Wieters and the de Meaux lab for testing and feedback on the browser, and all members of the Stetter lab for discussion and suggestions. We acknowledge the support of the Deutsche Forschungsgemeinschaft under Germany's Excellence Strategy – EXC-2048/1 – Project ID 390686111 to MGS.

4 Overarching Discussion

Understanding the historical context of crop domestication, along with the genetic regulation of key domestication traits, is crucial in light of rapidly changing environmental conditions that pose a threat to crop systems. The process of crop domestication has been carried out in various regions worldwide, resulting in crops with varying degrees of completeness. This variability can lead to less efficient crops that lack important domestication traits, resulting in lower yields and limited adaptability. Although crops with similar uses exhibit comparable domestication traits (Meyer et al., 2012), their path during domestication was likely as diverse as the species that have been cultivated (Stetter, 2020). The evolution of crop populations is the result of various processes, including selection, demographic changes, and gene flow. While selection and demographic changes have been extensively studied, the role of gene flow and its impact on the fitness of crops has only received attention in recent years. Despite progress in this field, the conceptual assumption of crop domestication as a linear process from one wild ancestor is still very present. Previous studies in amaranth and other species have been challenging that view (Stetter et al., 2020). In this thesis, I give further support to the hypothesis that current amaranth domesticated populations result from a complex relationship of factors and selective pressures across different locations and times.

The continuous exchanges between populations are prevalent, with signals of post-domestication introgression between the different crops, even at great geographical distances (Figure 2 from Section 2), signaling the ongoing evolution of the species, supporting previous reports of gene flow playing a significant role during domestication (Janzen et al., 2019). Introgression can have a range of effects on their genomes. On the one hand, the introgression of genome parts from wild species can lead to the homogenization of genomes (Slatkin, 1985). This homogenization can help to explain the proximity in morphology across the samples that led to difficulties in taxonomic classifications, which could be clarified using genomic resources (Figure S1). Moreover, adaptive gene flow from wild relatives into crops has been previously associated with environmental adaptation in crops. For example, in maize, the introgression of wild relative *Zea mays spp. mexicana* has been associated with maize's adaptation to highland conditions and colder climates (Wang et al., 2017). Although we cannot associate gene flow from wild relatives with positive selection, we found decreased genetic load in regions that were introgressed from wild relatives. This could be due to the higher effective population size and greater genetic diversity of wild relatives. The wild ancestor *A. hybridus* showed a lower genetic load than the domesticated varieties.

On the other hand, exchanges between wild and crops can introduce necessary genetic diversity to species that went through strong reproductive isolation (Ellstrand et al., 1999). During the process of

domestication, fewer individuals with desirable traits are kept, reducing the available genetic pool and, consequently, the genetic diversity (Figure 9). The results in Section 3 support this hypothesis, where domesticated populations had a lower nucleotide diversity than their wild ancestor *A. hybridus*. Further, Tajima's D was higher for the crop populations, which is an indicator of a domestication bottleneck. I could also demonstrate that domesticated grain amaranth has a higher deleterious mutation load. This is in line with the previous observation that a bottleneck can result in an accumulation of deleterious alleles and, therefore, a higher mutation load (Lu et al., 2006). Further, we demonstrated that introgression from wild populations could help to reduce the genetic burden of deleterious alleles, as populations with gene flow from wild amaranth saw a reduction of their genetic load. This beneficial introgression can help sustain the new crop species, providing genetic diversity that might prove essential for adaptation to the local conditions.

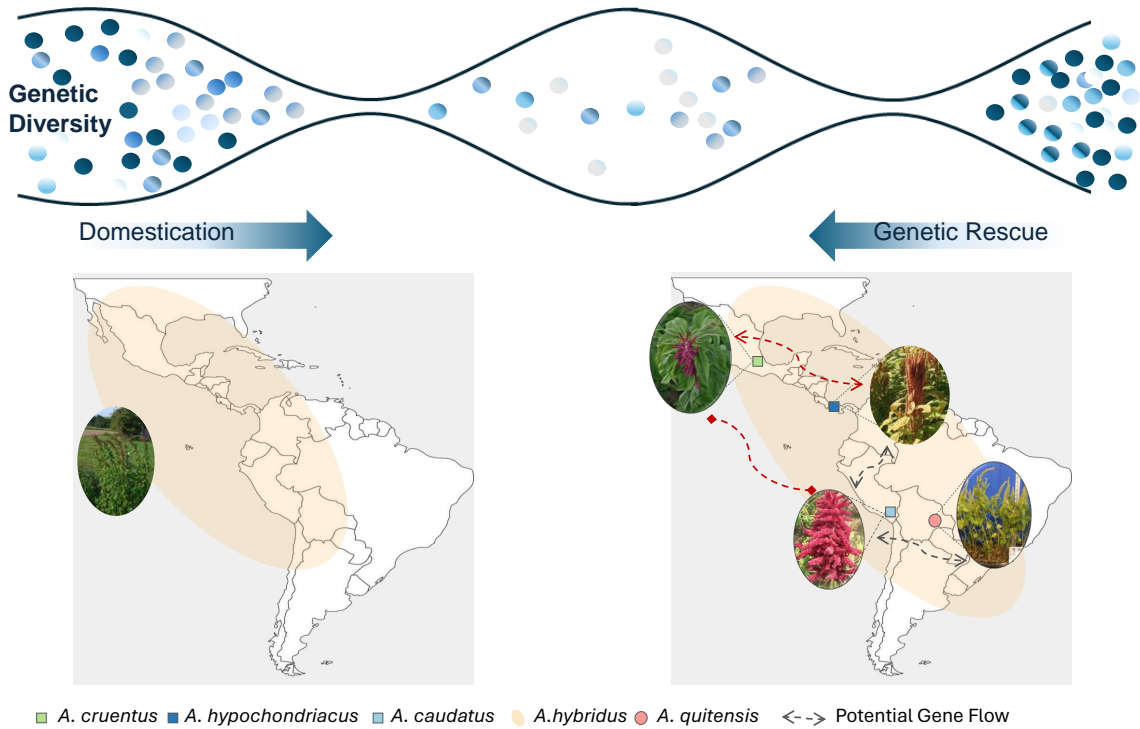


Figure 9: The contribution of multiple species to their genomes contributed to their genetic diversity, as we observe on the mosaic of their genome ancestries. Further, we demonstrated the dynamics of those relationships, with partial or very strong incompatibilities on the crosses (marked in red). Finally, the development of PopAmaranth offered an accessible tool for observing the summaries of the relationships between those populations.

However, introgressed blocks are not distributed uniformly across the genome. This mosaic and diversity could be observed in the local ancestral inference (Figure 3). Here, I could observe the contribution of the multiple species in the crop-wild system to the current population's genome. The differentiation presented in PopAmaranth supports the different genomic landscapes scenario; differentiation is higher between crops than between crops and the wild ancestor *A. hybridus*. Research into genomic data has confirmed our findings that genetic variation is not evenly distributed throughout the genome (Harris and Nielsen, 2016). It has been observed that some regions display noticeable differences between populations, indicating the intricate nature of genetic variation. This complexity may be attributed to various factors such as genetic drift, natural selection, and gene flow. Speciation leads to highly heterogeneous genetic variation along the genome, (Lohmueller et al., 2009; Lawson et al., 2012), which results in reproductive barriers with distinct genomic landscapes (Ravinet et al., 2017). *A. cruentus* has the highest differentiation and was also the more homogeneous population in the local ancestry inference tests. In contrast to the other amaranth species in the studies, *A. cruentus* has 17 chromosomes instead of 16 (Ma et al., 2021). While the difference in chromosome sizes alone does not explain the current incompatibility, as *A. cruentus* (Mesoamerica) was able to hybridize with some of the *A. hypochondriacus* accessions (also located in Mesoamerica), the crossing with *A. caudatus* (Southern America) result in unviable, necrotic pairs, in accordance with previous observations (Gupta and Gudu, 1991). Nonetheless, I could still observe signals of gene flow between *A. cruentus* and *A. caudatus* in current populations. Given their strong current incompatibilities, it is conceivable that during the long history of amaranth domestication, those exchanges were possible in the past and became more unfeasible as reproductive isolation evolved. The most likely explanation is the presence of a multi-locus Dobzhansky-Muller incompatibility (Muller, 1942), associated with neutrally evolved reproductive barrier for the more geographically distant populations. The current incomplete reproductive barrier between *A. hypochondriacus* and *A. cruentus* may provide an opportunity for understanding the progression of reproductive isolation during domestication (Tenaillon et al., 2023). In contrast, *A. caudatus* and *A. hypochondriacus* have demonstrated complete compatibility, making them promising candidates for heterotic pools for potential future amaranth breeding programs. Suppose the hybrid incompatibility arose at the advent of amaranth domestication, gene flow between *A. cruentus* and *A. caudatus* was likely beneficial, given the high fitness disadvantage of F₁ hybrids. We found a low but significant number of introgressed selective sweeps that might represent such beneficial gene flow between amaranth crop species that warrant future investigation.

In this work, I could also shed some more light on the very complex history of South American amaranth populations. *A. quitensis* has strong admixture with *A. caudatus* and high Tajima's D, indicating a bottleneck, agreeing with Stetter et al. (2017a, 2020). There is the possibility that *A. quitensis* is an intermediary species between *A. caudatus* and *A. hybridus*, but we cannot exclude it as a possible

feralized species from *A. caudatus*. Understanding these relationships would benefit from further studies.

Overall, the genetic makeup of modern grain amaranth populations is a complex result of various evolutionary processes, including gene flow, genetic incompatibilities, domestication bottlenecks, and the influence of wild relatives. The relationships between species are rather more complex than linear, with exchanges between populations despite large geographic distances and the reintroduction of genetic material that was potentially lost during speciation. Despite observed genetic incompatibilities and high genetic differentiation between crop species, strong signals of gene flow between grain amaranths and between crops and their wild relatives have been found. Recurrent gene flow from wild relatives into crops might have allowed evolutionary rescue, counteracting the loss of diversity but likely hindering the fixation of domestication traits, leading to the incomplete domestication syndrome observed today for grain amaranth. Further research using tools like the PopAmaranth genome browser and other accessible platforms for genomic and evolutionary data will help to better understand the complex relationships between crop species and their wild relatives, ultimately contributing to the development of more effective strategies for crop improvement and domestication.

5 Concluding Remarks

In this thesis, an in-depth investigation of gene flow in amaranth has provided valuable insights into its evolutionary history, particularly the continuous domestication process. The findings demonstrate that the domestication of crops, particularly amaranth, is characterized by a continuum of exchanges between wild and cultivated species, resulting in a complex genomic mosaic in current populations.

One fundamental discovery is the evidence of evolutionary rescue through gene flow between wild and domesticated populations, which effectively alleviates the genetic load and enhances the species' survival prospects. The study also reveals the heterogeneity of shared ancestry and gene flow across the genome and among different populations. These continuous exchanges have led to the formation of a mosaic pattern, challenging the notion of a linear evolution in amaranth.

Moreover, this research significantly advances our understanding of crop domestication, particularly in amaranth populations, highlighting the importance of genetic diversity and ongoing interactions between wild and domesticated populations for species preservation. These insights have broad implications for the conservation and sustainable use of crop genetic resources.

Additionally, the development of population genetic genome browsers, such as PopAmaranth, has greatly facilitated the analysis of population-scale summary statistics in various species. PopAmaranth offers a user-friendly platform that integrates features such as selection signals, gene annotation, and variant calls, enabling researchers to explore evolutionary patterns and identify convergent selection across taxa. The browser's utility extends beyond amaranth, showcasing its effectiveness in identifying selection signals in candidate genes across diverse species.

In summary, this thesis has provided new perspectives on the continuous domestication process, exemplified by amaranth. A comprehensive understanding of amaranth's evolutionary history has been attained by gene flow examination. The findings emphasize the crucial role of genetic diversity and ongoing interactions between wild and domesticated populations for species survival and adaptation. Furthermore, the development of population genetic genome browsers, such as PopAmaranth, has the potential to empower researchers to explore population-scale summary statistics and uncover selection signals across taxa, expanding our knowledge in this field.

6 Bibliography

- Shahal Abbo, Ruth Pinhasi van Oss, Avi Gopher, Yehoshua Saranga, Itai Ofner, and Zvi Peleg. Plant domestication versus crop evolution: a conceptual framework for cereals and grain legumes. *Trends in Plant Science*, 19(6):351–360, 2014.
- Stephanie M Aguillon, Tristram O Dodge, Gabriel A Preising, and Molly Schumer. Introgression. *Current Biology*, 32(16):R865–R868, 2022.
- Nikolaos Alachiotis and Pavlos Pavlidis. RAiSD detects positive selection based on multiple signatures of a selective sweep and SNP vectors. *Communications Biology*, 1(1):1–11, 2018.
- Carlos Alonso-Blanco, Jorge Andrade, Claude Becker, Felix Bemm, Joy Bergelson, Karsten M Borgwardt, Jun Cao, Eunyoung Chae, Todd M Dezwaan, Wei Ding, et al. 1,135 genomes reveal the global pattern of polymorphism in *Arabidopsis thaliana*. *Cell*, 166(2):481–491, 2016.
- Stephen F Altschul, Warren Gish, Webb Miller, Eugene W Myers, and David J Lipman. Basic local alignment search tool. *Journal of Molecular Biology*, 215(3):403–410, 1990.
- Hannah J Appiah-Madson, Eric B Knox, Christina M Caruso, and Andrea L Case. Do genetic drift and gene flow affect the geographic distribution of female plants in gynodioecious *Lobelia siphilitica*? *Plants*, 11(6):825, 2022.
- Jean Armengaud, Judith Trapp, Olivier Pible, Olivier Geffard, Arnaud Chaumot, and Erica M Hartmann. Non-model organisms, a species endangered by proteogenomics. *Journal of Proteomics*, 105:5–18, 2014.
- JS Aswal, BS Bisht, Rajendra Dobhal, and DP Uniyal. Historical journey with amaranth. *Asian Agri-History*, 20(3), 2016.
- Giridhar Athrey, Nikolas Faust, Anne-Sophie Charlotte Hieke, and I Lehr Brisbin. Effective population sizes and adaptive genetic variation in a captive bird population. *PeerJ*, 6:e5803, 2018.
- Eric J Baack and Loren H Rieseberg. A genomic view of introgression and hybrid speciation. *Journal Current Opinion in Genetics & Development of Proteomics*, 17(6):513–518, 2007.
- Connor Bernard, Aldo Compagnoni, and Roberto Salguero-Gómez. Testing finch’s hypothesis: The role of organismal modularity on the escape from actuarial senescence. *Functional Ecology*, 34(1):88–106, 2020.

- Giorgio Bertorelle, Francesca Raffini, Mirte Bosse, Chiara Bortoluzzi, Alessio Iannucci, Emiliano Trucchi, Hernán E Morales, and Cock Van Oosterhout. Genetic load: genomic estimates and applications in non-model animals. *Nature Reviews Genetics*, 23(8):492–503, 2022.
- Mathieu Blanchette, W James Kent, Cathy Riemer, Laura Elnitski, Arian FA Smit, Krishna M Roskin, Robert Baertsch, Kate Rosenbloom, Hiram Clawson, Eric D Green, et al. Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Research*, 14(4):708–715, 2004.
- Dan Bolser, Daniel M Staines, Emily Pritchard, and Paul Kersey. Ensembl plants: integrating tools for visualizing, mining, and analyzing plant genomics data. In *Plant Bioinformatics*, pages 115–140. Springer, 2016.
- Brian L. Browning, Xiaowen Tian, Ying Zhou, and Sharon R. Browning. Fast two-stage phasing of large-scale sequence data. *The American Journal of Human Genetics*, 108(10):1880–1890, 2021. ISSN 0002-9297. doi: 10.1016/j.ajhg.2021.08.005.
- KB Budde, Santiago C González-Martínez, Olivier J Hardy, and Myriam Heuertz. The ancient tropical rainforest tree *Symphonia globulifera* l. f.(clusiaceae) was not restricted to postulated pleistocene refugia in atlantic equatorial africa. *Heredity*, 111(1):66–76, 2013.
- Molly K Burke, Gianni Liti, and Anthony D Long. Standing genetic variation drives repeatable experimental evolution in outcrossing populations of *saccharomyces cerevisiae*. *Molecular Biology and Evolution*, 31(12):3228–3239, 2014.
- Angelo Canty and Brian Ripley. Package ‘boot’. *Bootstrap Functions. CRAN R Proj*, 2017.
- Stephanie M Carlson, Curry J Cunningham, and Peter AH Westley. Evolutionary rescue in a changing world. *Trends in Ecology & Evolution*, 29(9):521–530, 2014.
- Sònia Casillas, Roger Mulet, Pablo Villegas-Mirón, Sergi Hervas, Esteve Sanz, Daniel Velasco, Jaume Bertranpetit, Hafid Laayouni, and Antonio Barbadilla. PopHuman: the human population genomics browser. *Nucleic Acids Research*, 46(D1):D1003–D1010, 2018.
- Andrea P Castellanos-Arévalo, Andrés A Estrada-Luna, José L Cabrera-Ponce, Eliana Valencia-Lozano, Humberto Herrera-Ubaldo, Stefan de Folter, Alejandro Blanco-Labra, and John P Délano-Frier. *Agrobacterium* rhizogenes-mediated transformation of grain (*Amaranthus hypochondriacus*) and leafy (*A. hybridus*) amaranths. *Plant Cell Reports*, 2020.

- Mathilde Causse, Nelly Desplat, Laura Pascual, Marie-Christine Le Paslier, Christopher Sauvage, Guillaume Bauchet, Aurélie Bérard, Rémi Bounon, Maria Tchoumakov, Dominique Brunel, et al. Whole genome resequencing in tomato reveals variation associated with introgression and breeding events. *BMC Genomics*, 14(1):1–14, 2013.
- Igor Cesarino, Raffaele Dello Ioio, Gwendolyn K Kirschner, Michael S Ogden, Kelsey L Picard, Madlen I Rast-Somssich, and Marc Somssich. Plant science’s next top models. *Annals of Botany*, 126(1):1–23, 2020.
- B Charlesworth and D Charlesworth. Population genetics from 1966 to 2016. *Heredity*, 118(1):2–9, 2017. ISSN 0018-067X. doi: 10.1038/hdy.2016.55.
- Brian Charlesworth. Effective population size and patterns of molecular evolution and variation. *Nature Reviews Genetics*, 10(3):195–205, 2009.
- Guillaume Chomicki, Hanno Schaefer, and Susanne S Renner. Origin and domestication of Cucurbitaceae crops: insights from phylogenies, genomics and archaeology. *New Phytologist*, 226(5):1240–1255, 2020.
- Peter Civián, Konstantina Drosou, David Armisen-Gimenez, Wandrille Duchemin, Jérôme Salse, and Terence A Brown. Episodes of gene flow and selection during the evolutionary history of domesticated barley. *BMC Genomics*, 22:1–17, 2021.
- JW Clouse, D Adhikary, JT Page, T Ramaraj, MK Deyholos, JA Udall, DJ Fairbanks, EN Jellen, and PJ Maughan. The amaranth genome: genome, transcriptome, and physical map assembly. *The Plant Genome*, 9(1):1–14, 2016.
- Mihai Costea, Andrew Sanders, and Giles Waines. Preliminary results toward a revision of the *Amaranthus hybridus* species complex (*Amaranthaceae*). *SIDA, Contributions to Botany*, pages 931–974, 2001.
- Erika Crispo. Modifying effects of phenotypic plasticity on interactions among natural selection, adaptation and gene flow. *Journal of Evolutionary Biology*, 21(6):1460–1469, 2008.
- Lisa G Crozier, AP Hendry, Peter W Lawson, TP Quinn, NJ Mantua, J Battin, RG Shaw, and RB3352429 Huey. Potential responses to climate change in organisms with complex life histories: evolution and plasticity in pacific salmon. *Evolutionary Applications*, 1(2):252–270, 2008.
- Mitchell B Cruzan. Testing wright’s intermediate population size hypothesis—when genetic drift is a good thing. *bioRxiv*, pages 2022–09, 2022.

- Petr Danecek, Adam Auton, Goncalo Abecasis, Cornelis A Albers, Eric Banks, Mark A DePristo, Robert E Handsaker, Gerton Lunter, Gabor T Marth, Stephen T Sherry, et al. The variant call format and VCFtools. *Bioinformatics*, 27(15):2156–2158, 2011.
- Sudhansu Dash, Jacqueline D Campbell, Ethalinda KS Cannon, Alan M Cleary, Wei Huang, Scott R Kalberer, Vijay Karingula, Alex G Rice, Jugpreet Singh, Pooja E Umale, et al. Legume information system (LegumeInfo.org): a key component of a set of federated data resources for the legume family. *Nucleic Acids Research*, 44(D1):D1181–D1188, 2016.
- Eugene V Davydov, David L Goode, Marina Sirota, Gregory M Cooper, Arend Sidow, and Serafim Batzoglou. Identifying a high fraction of the human genome to be under selective constraint using gerp++. *PLoS Computational Biology*, 6(12):e1001025, 2010.
- Pierre De Wit, Melissa H Pespeni, Jason T Ladner, Daniel J Barshis, François Seneca, Hannah Jaris, Nina Overgaard Therkildsen, Megan Morikawa, and Stephen R Palumbi. The simple fool’s guide to population genomics via rna-seq: an introduction to high-throughput sequencing data analysis. *Molecular Ecology Resources*, 12(6):1058–1067, 2012.
- Hannes Dempewolf, Kathryn A Hodgins, Sonja E Rummell, Norman C Ellstrand, and Loren H Rieseberg. Reproductive isolation during domestication. *The Plant Cell*, 24(7):2710–2717, 2012.
- Sheila O Denn and W John MacMullen. The ambiguous bioinformatics domain: A conceptual map of information science applications for molecular biology. *Proceedings of the American Society for Information Science and Technology*, 39(1):556–558, 2002.
- Ya-Mei Ding, Yu Cao, Wei-Ping Zhang, Jun Chen, Jie Liu, Pan Li, Susanne S Renner, Da-Yong Zhang, and Wei-Ning Bai. Population-genomic analyses reveal bottlenecks and asymmetric introgression from Persian into iron walnut during domestication. *Genome Biology*, 23(1):1–18, 2022.
- John Doebley. Unfallen grains: how ancient farmers turned weeds into crops. *Science*, 312(5778):1318–1319, 2006.
- John F Doebley, Brandon S Gaut, and Bruce D Smith. The molecular genetics of crop domestication. *Cell*, 127(7):1309–1321, 2006.
- Julien Y Dutheil, Sylvain Gaillard, and Eva H Stukenbrock. Maffilter: a highly flexible and extensible multiple genome alignment files processor. *BMC Genomics*, 15:1–10, 2014.
- Sean R Eddy. “antedisciplinary” science. *PLoS Computational Biology*, 1(1):e6, 2005.

- Greizerstein EJ and Lidia Poggio. Karyological studies in grain amaranths. *Cytologia*, 59(1):25–30, 1994.
- Norman C Ellstrand. Is gene flow the most important evolutionary force in plants? *American Journal of Botany*, 101(5):737–753, 2014.
- Norman C Ellstrand, Honor C Prentice, and James F Hancock. Gene flow and introgression from domesticated plants into their wild relatives. *Annual Review of Ecology and Systematics*, 30(1):539–563, 1999.
- Laurent Excoffier, Isabelle Dupanloup, Emilia Huerta-Sánchez, Vitor C Sousa, and Matthieu Foll. Robust demographic inference from genomic and SNP data. *PLoS Genetics*, 9(10):e1003905, 2013.
- Adam Eyre-Walker, Rebecca L Gaut, Holly Hilton, Dawn L Feldman, and Brandon S Gaut. Investigation of the bottleneck leading to the domestication of maize. *Proceedings of the National Academy of Sciences*, 95(8):4441–4446, 1998.
- Jeffrey L Feder, Scott P Egan, and Patrik Nosil. The genomics of speciation-with-gene-flow. *Trends in Genetics*, 28(7):342–350, 2012.
- Noe Fernandez-Pozo, Naama Menda, Jeremy D Edwards, Surya Saha, Isaak Y Teclé, Susan R Strickler, Aureliano Bombarely, Thomas Fisher-York, Anuradha Pujar, Hartmut Foerster, et al. The Sol Genomics Network (SGN)—from genotype to phenotype to breeding. *Nucleic Acids Research*, 43(D1):D1036–D1041, 2015.
- Alisdair R Fernie and Jianbing Yan. De novo domestication: an alternative route toward new crops for the future. *Molecular Plant*, 12(5):615–631, 2019.
- Laurent AF Frantz, Joshua G Schraiber, Ole Madsen, Hendrik-Jan Megens, Alex Cagan, Mirte Bosse, Yogesh Paudel, Richard PMA Crooijmans, Greger Larson, and Martien AM Groenen. Evidence of long-term gene flow and selection during domestication from analyses of eurasian wild and domestic pig genomes. *Nature Genetics*, 47(10):1141–1148, 2015.
- Brandon S Gaut, Concepción M Díez, and Peter L Morrell. Genomics and the contrasting dynamics of annual and perennial domestication. *Trends in Genetics*, 31(12):709–719, 2015.
- Brandon S Gaut, Danelle K Seymour, Qingpo Liu, and Yongfeng Zhou. Demography and its effects on genomic variation in crop domestication. *Nature Plants*, 4(8):512–520, 2018.

- Eben Gering, Darren Inorvaia, Rie Henriksen, Jeffrey Conner, Thomas Getty, and Dominic Wright. Getting back to nature: feralization in animals and plants. *Trends in Ecology & Evolution*, 34(12): 1137–1151, 2019.
- Zachariah Gompert, Amy Springer, Megan Brady, Samridhi Chaturvedi, and Lauren K Lucas. Genomic time-series data show that gene flow maintains high genetic diversity despite substantial genetic drift in a butterfly species. *Molecular Ecology*, 30(20):4991–5008, 2021.
- José Gonçalves-Dias, Akanksha Singh, Corbinian Graf, and Markus G Stetter. Genetic incompatibilities and evolutionary rescue by wild relatives shaped grain amaranth domestication. *Molecular Biology and Evolution*, 40(8):msad177, 2023.
- José Gonçalves-Dias and Markus G Stetter. PopAmaranth: A population genetic genome browser for grain amaranths and their wild relatives. *G3: Genes, Genomes, Genetics*, 2021. doi: 10.1101/2020.12.09.415331.
- David M Goodstein, Shengqiang Shu, Russell Howson, Rochak Neupane, Richard D Hayes, Joni Fazo, Therese Mitros, William Dirks, Uffe Hellsten, Nicholas Putnam, et al. Phytozome: a comparative platform for green plant genomics. *Nucleic Acids Research*, 40(D1):D1178–D1186, 2012.
- Toni I Gossmann, Megan Woolfit, and Adam Eyre-Walker. Quantifying the variation in the effective population size within a genome. *Genetics*, 189(4):1389–1402, 2011.
- VK Gupta and S Gudu. Interspecific hybrids and possible phylogenetic relations in grain amaranths. *Euphytica*, 52(1):33–38, 1991.
- Karl Hammer. Das domestikationssyndrom. *Kulturpflanze*, 32:11–34, 1984.
- Kelley Harris and Rasmus Nielsen. The genetic cost of Neanderthal introgression. *Genetics*, 203(2): 881–891, 2016.
- Holly Hilton and Brandon S Gaut. Speciation and domestication in maize and its wild relatives: evidence from the globulin-1 gene. *Genetics*, 150(2):863–872, 1998.
- Kaichi Huang, Mojtaba Jahani, Jérôme Gouzy, Alexandra Legendre, Sébastien Carrere, José Miguel Lázaro-Guevara, Eric Gerardo González Segovia, Marco Todesco, Baptiste Mayjonade, Nathalie Rodde, et al. The genomics of linkage drag in inbred lines of sunflower. *Proceedings of the National Academy of Sciences*, 120(14):e2205783119, 2023.

- Xuehui Huang, Nori Kurata, Zi-Xuan Wang, Ahong Wang, Qiang Zhao, Yan Zhao, Kunyan Liu, Hengyun Lu, Wenjun Li, Yunli Guo, et al. A map of rice genome variation reveals the origin of cultivated rice. *Nature*, 490(7421):497–501, 2012.
- Tim Hubbard, Daniel Barker, Ewan Birney, Graham Cameron, Yuan Chen, L Clark, Tony Cox, J Cuff, Val Curwen, Thomas Down, et al. The ensembl genome database project. *Nucleic Acids Research*, 30(1):38–41, 2002.
- Jaime Huerta-Cepas, François Serra, and Peer Bork. ETE 3: Reconstruction, analysis, and visualization of phylogenomic data. *Molecular Biology and Evolution*, 33(6):1635–1638, 2016.
- JA Huerta-Ocampo, MF León-Galván, LB Ortega-Cruz, A Barrera-Pacheco, A De León-Rodríguez, G Mendoza-Hernández, and AP Barba de la Rosa. Water stress induces up-regulation of DOF1 and MIF1 transcription factors and down-regulation of proteins involved in secondary metabolism in amaranth roots (*Amaranthus hypochondriacus* L.). *Plant Biology*, 13(3):472–482, 2011.
- Matthew B Hufford, Xun Xu, Joost Van Heerwaarden, Tanja Pyhäjärvi, Jer-Ming Chia, Reed A Cartwright, Robert J Elshire, Jeffrey C Glaubitz, Kate E Guill, Shawn M Kaeppler, et al. Comparative population genomics of maize domestication and improvement. *Nature Genetics*, 44(7):808, 2012.
- Matthew B. Hufford, Pesach Lubinsky, Tanja Pyhäjärvi, Michael T. Devengenzo, Norman C. Ellstrand, and Jeffrey Ross-Ibarra. The genomic signature of crop-wild introgression in maize. *PLoS Genetics*, 9(5), 2013. ISSN 1553-7390. doi: 10.1371/journal.pgen.1003477.
- Arabidopsis Genome Initiative. Analysis of the genome sequence of the flowering plant arabidopsis thaliana. *Nature*, 408(6814):796–815, 2000.
- Hideki Innan and Yuseob Kim. Pattern of polymorphism after strong artificial selection in a domestication event. *Proceedings of the National Academy of Sciences*, 101(29):10667–10672, 2004.
- Garrett M. Janzen, Li Wang, and Matthew B. Hufford. The extent of adaptive wild introgression in crops. *New Phytologist*, 221(3):1279–1288, 2019. ISSN 1469-8137. doi: 10.1111/nph.15457. URL <https://onlinelibrary.wiley.com/doi/abs/10.1111/nph.15457>.
- David E Jarvis, Yung Shwen Ho, Damien J Lightfoot, Sandra M Schmöckel, Bo Li, Theo JA Borm, Hajime Ohyanagi, Katsuhiko Mineta, Craig T Michell, Noha Saber, et al. The genome of *Chenopodium quinoa*. *Nature*, 542(7641):307, 2017.

- Jingjing Jin, Jun Liu, Huan Wang, Limsoon Wong, and Nam-Hai Chua. PLncDB: plant long non-coding RNA database. *Bioinformatics*, 29(8):1068–1071, 2013.
- Dinesh C Joshi, Salej Sood, Rajashekara Hosahatti, Lakshmi Kant, A Pattanayak, Anil Kumar, Dinesh Yadav, and Markus G Stetter. From zero to hero: the past, present and future of grain amaranth breeding. *Theoretical and Applied Genetics*, 131(9):1807–1823, 2018.
- Donna Karolchik, Robert Baertsch, Mark Diekhans, Terrence S Furey, Angie Hinrichs, YT Lu, Krishna M Roskin, Matt Schwartz, Charles W Sugnet, Daryl J Thomas, et al. The ucsc genome browser database. *Nucleic Acids Research*, 31(1):51–54, 2003.
- W. J. Kent, A. S. Zweig, G. Barber, A. S. Hinrichs, and D. Karolchik. BigWig and BigBed: enabling browsing of large distributed datasets. *Bioinformatics*, 26(17):2204–2207, 2010. ISSN 1367-4803. doi: 10.1093/bioinformatics/btq351.
- Szymon M Kielbasa, Raymond Wan, Kengo Sato, Paul Horton, and Martin C Frith. Adaptive seeds tame genomic sequence comparison. *Genome Research*, 21(3):487–493, 2011.
- Kiel D Kietlinski, Felix Jimenez, Eric N Jellen, Peter J Maughan, Scott M Smith, and Donald B Pratt. Relationships between the weedy *Amaranthus hybridus* (*Amaranthaceae*) and the grain amaranths. *Crop Science*, 54(1):220–228, 2014.
- Myung-Shin Kim, Roberto Lozano, Ji Hong Kim, Dong Nyuk Bae, Sang-Tae Kim, Jung-Ho Park, Man Soo Choi, Jaehyun Kim, Hyun-Choong Ok, Soo-Kwon Park, et al. The patterns of deleterious mutations during the domestication of soybean. *Nature Communications*, 12(1):97, 2021.
- Thorfinn Sand Korneliussen, Anders Albrechtsen, and Rasmus Nielsen. ANGSD: analysis of next generation sequencing data. *BMC bioinformatics*, 15(1):356, 2014a.
- Thorfinn Sand Korneliussen, Anders Albrechtsen, and Rasmus Nielsen. ANGSD: Analysis of Next Generation Sequencing Data. *BMC Bioinformatics*, 15(1):356, 2014b. ISSN 1471-2105. doi: 10.1186/s12859-014-0356-4. URL <https://doi.org/10.1186/s12859-014-0356-4>.
- Krzysztof M Kozak, Mathieu Joron, W Owen McMillan, and Chris D Jiggins. Rampant genome-wide admixture across the *Heliconius* radiation. *Genome Biology and Evolution*, 13(7):evab099, 2021.
- Liz Koziol, Loren H Rieseberg, Nolan Kane, and James D Bever. Reduced drought tolerance during domestication and the evolution of weediness results from tolerance–growth trade-offs. *Evolution: International Journal of Organic Evolution*, 66(12):3803–3814, 2012.

- Julia M Kreiner, Darci Ann Giacomini, Felix Bemm, Bridgit Waithaka, Julian Regalado, Christa Lanz, Julia Hildebrandt, Peter H Sikkema, Patrick J Tranel, Detlef Weigel, et al. Multiple modes of convergent adaptation in the spread of glyphosate-resistant *amaranthus tuberculatus*. *Proceedings of the National Academy of Sciences*, 116(42):21076–21084, 2019.
- Vivek Krishnakumar, Matthew R Hanlon, Sergio Contrino, Erik S Ferlanti, Svetlana Karamycheva, Maria Kim, Benjamin D Rosen, Chia-Yi Cheng, Walter Moreira, Stephen A Mock, et al. Araport: the Arabidopsis information portal. *Nucleic Acids Research*, 43(D1):D1003–D1009, 2015.
- Toru Kudo, Masaaki Kobayashi, Shin Terashima, Minami Katayama, Soichi Ozaki, Maasa Kanno, Misa Saito, Koji Yokoyama, Hajime Ohyanagi, Koh Aoki, et al. TOMATOMICS: a web database for integrated omics information in tomato. *Plant and Cell Physiology*, 58(1):e8–e8, 2017.
- Brijesh Kumar and Purva Bhalothia. Orphan crops for future food security. *Journal of biosciences*, 45: 1–8, 2020.
- Hon-Ming Lam, Xun Xu, Xin Liu, Wenbin Chen, Guohua Yang, Fuk-Ling Wong, Man-Wah Li, Weiming He, Nan Qin, Bo Wang, et al. Resequencing of 31 wild and cultivated soybean genomes identifies patterns of genetic diversity and selection. *Nature Genetics*, 42(12):1053–1059, 2010.
- Carolyn J Lawrence, Qunfeng Dong, Mary L Polacco, Trent E Seigfried, and Volker Brendel. MaizeGDB, the community database for maize genetics and genomics. *Nucleic Acids Research*, 32(suppl_1):D393–D397, 2004.
- Daniel Lawson, Garrett Hellenthal, Simon Myers, and Daniel Falush. Inference of population structure using dense haplotype data. *PLoS Genetics*, 8(1), 2012. ISSN 1553-7390. doi: 10.1371/journal.pgen.1002453.
- Zachary H Lemmon, Nathan T Reem, Justin Dalrymple, Sebastian Soyk, Kerry E Swartwood, Daniel Rodriguez-Leal, Joyce Van Eck, and Zachary B Lippman. Rapid improvement of domestication traits in an orphan crop by genome editing. *Nature Plants*, 4(10):766–770, 2018.
- Sebastian Lequime, Albin Fontaine, Meriadeg Ar Gouilh, Isabelle Moltini-Conclois, and Louis Lambrechts. Genetic drift, purifying selection and vector genotype shape dengue virus intra-host genetic diversity in mosquitoes. *PLoS genetics*, 12(6):e1006111, 2016.
- Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, and Richard Durbin. The sequence alignment/map format and SAMtools. *Bioinformatics*, 25(16):2078–2079, 2009.

- DJ Lightfoot, David Erwin Jarvis, T Ramaraj, R Lee, EN Jellen, and PJ Maughan. Single-molecule sequencing and Hi-C-based proximity-guided assembly of amaranth (*Amaranthus hypochondriacus*) chromosomes provide insights into genome evolution. *BMC biology*, 15(1):74, 2017a.
- DJ Lightfoot, DE Jarvis, T Ramaraj, R Lee, EN Jellen, and PJ Maughan. Single-molecule sequencing and Hi-C-based proximity-guided assembly of amaranth (*Amaranthus hypochondriacus*) chromosomes provide insights into genome evolution. *BMC Biology*, 15(1):74, 2017b. doi: 10.1186/s12915-017-0412-4.
- Shuo Liu, Amandine Cornille, Stéphane Decroocq, David Tricon, Aurélie Chague, Jean-Philippe Eyquard, Wei-Sheng Liu, Tatiana Giraud, and Véronique Decroocq. The complex evolutionary history of apricots: Species divergence, gene flow and multiple domestication events. *Molecular Ecology*, 28(24): 5299–5314, 2019.
- Kirk E Lohmueller, Carlos D Bustamante, and Andrew G Clark. Methods for human demographic inference using haplotype patterns from genomewide single-nucleotide polymorphism data. *Genetics*, 182(1):217–231, 2009.
- Xosé López-Goldar and Anurag A Agrawal. Ecological interactions, environmental gradients, and gene flow in local adaptation. *Trends in Plant Science*, 26(8):796–809, 2021.
- Roberto Lozano, Elodie Gazave, Jhonathan PR Dos Santos, Markus G Stetter, Ravi Valluru, Nonoy Bandillo, Samuel B Fernandes, Patrick J Brown, Nadia Shakoor, Todd C Mockler, et al. Comparative evolutionary genetics of deleterious load in sorghum and maize. *Nature Plants*, 7(1):17–24, 2021.
- Jian Lu, Tian Tang, Hua Tang, Jianzi Huang, Suhua Shi, and Chung-I. Wu. The accumulation of deleterious mutations in rice genomes: A hypothesis on the cost of domestication. *Trends in Genetics*, 22(3):126–131, 2006. ISSN 0168-9525. doi: 10.1016/j.tig.2006.01.004.
- Gordon Luikart, Marty Kardos, Brian K Hand, Om P Rajora, Sally N Aitken, and Paul A Hohenlohe. Population genomics: advancing understanding of nature. *Population genomics: concepts, approaches and applications*, pages 3–79, 2019.
- M-C Luo, Z-L Yang, FM You, T Kawahara, JG Waines, and J Dvorak. The structure of wild and domesticated emmer wheat populations, gene flow between them, and the site of emmer domestication. *Theoretical and Applied Genetics*, 114:947–959, 2007.
- Feng-Hua Lv, Yin-Hong Cao, Guang-Jian Liu, Ling-Yun Luo, Ran Lu, Ming-Jun Liu, Wen-Rong Li, Ping Zhou, Xin-Hua Wang, Min Shen, et al. Whole-genome resequencing of worldwide wild and domestic

- sheep elucidates genetic diversity, introgression, and agronomically important loci. *Molecular Biology and Evolution*, 39(2):msab353, 2022.
- Xiao Ma, Fabian E Vaistij, Yi Li, Willem S Jansen van Rensburg, Sarah Harvey, Michael W Bairu, Sonja L Venter, Sydney Mavengahama, Zemin Ning, Ian A Graham, et al. A chromosome-level *Amaranthus cruentus* genome assembly highlights gene family evolution and biosynthetic gene clusters that may underpin the nutritional value of this traditional crop. *The Plant Journal*, 107(2):613–628, 2021.
- Milan Malinsky, Michael Matschiner, and Hannes Svoldal. Dsuite - Fast D -statistics and related admixture evidence from VCF files. *Molecular Ecology Resources*, 21(2):584–595, 2021-02. ISSN 1755-098X, 1755-0998. doi: 10.1111/1755-0998.13265.
- Melanie A Mallory, Rozaura V Hall, Andrea R McNabb, Donald B Pratt, Eric N Jellen, and Peter J Maughan. Development and characterization of microsatellite markers for the grain amaranths. *Crop science*, 48(3):1098–1106, 2008.
- Locedie Mansueto, Roven Rommel Fuentes, Frances Nikki Borja, Jeffery Detras, Juan Miguel Abriol-Santos, Dmytro Chebotarov, Millicent Sanciangco, Kevin Palis, Dario Copetti, Alexandre Poliakov, et al. Rice SNP-seek database update: new SNPs, indels, and queries. *Nucleic Acids Research*, 45(D1):D1075–D1081, 2017.
- Simon H Martin and Steven M Van Belleghem. Exploring evolutionary relationships across the genome using topology weighting. *Genetics*, 206(1):429–438, 2017.
- Joanna Masel. Genetic drift. *Current Biology*, 21(20):R837–R838, 2011.
- Julio A Massange-Sanchez, Paola A Palmeros-Suarez, Eduardo Espitia-Rangel, Isaac Rodriguez-Arevalo, Lino Sanchez-Segura, Norma A Martinez-Gallardo, Fulgencio Alatorre-Cobos, Axel Tiessen, and John P Delano-Frier. Overexpression of grain amaranth (*Amaranthus hypochondriacus*) AhERF or AhDOF transcription factors in *Arabidopsis thaliana* increases water deficit-and salt-stress tolerance, respectively, via contrasting stress-amelioration mechanisms. *PLoS ONE*, 11(10):e0164280, 2016.
- Julio Armando Massange-Sanchez, Paola Andrea Palmeros-Suarez, Norma Angelica Martinez-Gallardo, Paula Andrea Castrillon-Arbelaez, Hamlet Aviles-Arnaut, Fulgencio Alatorre-Cobo, Axel Tiessen, and John Paul Délano-Frier. The novel and taxonomically restricted Ah24 gene from grain amaranth (*Amaranthus hypochondriacus*) has a dual role in development and defense. *Frontiers in Plant Science*, 6:602, 2015.

- S Mayes, FJ Massawe, PG Alderson, JA Roberts, SN Azam-Ali, and M Hermann. The potential for underutilized crops to improve security of food production. *Journal of Experimental Botany*, 63(3): 1075–1079, 2012.
- Jonas Meisner and Anders Albrechtsen. Inferring population structure and admixture proportions in low-depth NGS data. *Genetics*, 210(2):719–731, 2018.
- Rachel S Meyer, Ashley E DuVal, and Helen R Jensen. Patterns and processes in crop domestication: an historical review and quantitative analysis of 203 global food crops. *New Phytologist*, 196(1):29–48, 2012.
- Rachel S Meyer, Jae Young Choi, Michelle Sanches, Anne Plessis, Jonathan M Flowers, Junrey Amas, Katherine Dorph, Annie Barretto, Briana Gross, Dorian Q Fuller, et al. Domestication history and geographical adaptation inferred from a snp map of african rice. *Nature Genetics*, 48(9):1083–1088, 2016.
- Thomas Mitchell-Olds, John H Willis, and David B Goldstein. Which evolutionary processes influence natural genetic variation for phenotypic traits? *Nature Reviews Genetics*, 8(11):845–856, 2007.
- Luca Montana, Romolo Caniglia, Marco Galaverni, Elena Fabbri, Atidje Ahmed, Barbora Černá Bolfíková, Sylwia D Czarnomska, Ana Galov, Raquel Godinho, Maris Hindrikson, et al. Combining phylogenetic and demographic inferences to assess the origin of the genetic diversity in an isolated wolf population. *PLoS ONE*, 12(5):e0176560, 2017.
- Jacob S Montgomery, Darci Giacomini, Bridgit Waithaka, Christa Lanz, Brent P Murphy, Ruth Campe, Jens Lerchl, Andreas Landes, Fanny Gatzmann, Antoine Janssen, et al. Draft Genomes of *Amaranthus tuberculatus*, *Amaranthus hybridus*, and *Amaranthus palmeri*. *Genome Biology and Evolution*, 12(11): 1988–1993, 2020.
- Alejandra Moreno-Letelier, Jonás A Aguirre-Liguori, Daniel Piñero, Alejandra Vázquez-Lobo, and Luis E Eguiarte. The relevance of gene flow with wild relatives in understanding the domestication process. *Royal Society Open Science*, 7(4):191545, 2020.
- Agnieszka Mroczek. Phytochemistry and bioactivity of triterpene saponins from *Amaranthaceae* family. *Phytochemistry Reviews*, 14(4):577–605, 2015.
- Hermann J Muller. Isolating mechanisms, evolution, and temperature. In *Biol Symp*, volume 6, pages 71–125, 1942.

- Benoit Nabholz, Gautier Sarah, François Sabot, Manuel Ruiz, Hélène Adam, Sabine Nidelet, Alain Ghesquière, Sylvain Santoni, Jacques David, and Sylvain Glémin. Transcriptome population genomics reveals severe bottleneck and domestication cost in the African rice (*Oryza glaberrima*). *Molecular Ecology*, 23(9):2210–2227, 2014.
- Masatoshi Nei and Wen-Hsiung Li. Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proceedings of the National Academy of Sciences*, 76(10):5269–5273, 1979.
- Suzhen Niu, Qinfei Song, Hisashi Koiwa, Dahe Qiao, Degang Zhao, Zhengwu Chen, Xia Liu, and Xiaopeng Wen. Genetic diversity, linkage disequilibrium, and population structure analysis of the tea plant (*Camellia sinensis*) from an origin center, guizhou plateau, using genome-wide snps developed by genotyping-by-sequencing. *BMC Plant Biology*, 19(1):1–12, 2019.
- Irene Novo, Enrique Santiago, and Armando Caballero. The estimates of effective population size based on linkage disequilibrium are virtually unaffected by natural selection. *PLoS Genetics*, 18(1):e1009764, 2022.
- Wieslaw Oleszek, Marta Junkuszew, and Anna Stochmal. Determination and toxicity of saponins from *Amaranthus cruentus* seeds. *Journal of Agricultural and Food Chemistry*, 47(9):3685–3687, 1999.
- Ludovic Orlando, Robin Allaby, Pontus Skoglund, Clio Der Sarkissian, Philipp W Stockhammer, María C Ávila-Arcos, Qiaomei Fu, Johannes Krause, Eske Willerslev, Anne C Stone, et al. Ancient DNA analysis. *Nature Reviews Methods Primers*, 1(1):14, 2021.
- Katherine L Ostevik, Rose L Andrew, Sarah P Otto, and Loren H Rieseberg. Multiple reproductive barriers separate recently diverged sunflower ecotypes. *Evolution*, 70(10):2322–2335, 2016.
- Carolina Osuna-Mascaró, Rafael Rubio de Casas, José M Gómez, João Loureiro, Silvia Castro, Jacob B Landis, Robin Hopkins, and Francisco Perfectti. Hybridization and introgression are prevalent in southern european *Erysimum* (brassicaceae) species. *Annals of Botany*, 131(1):171–184, 2023.
- Anna Page, Jane Gibson, Rachel S Meyer, and Mark A Chapman. Eggplant domestication: pervasive gene flow, feralization, and transcriptomic divergence. *Molecular Biology and Evolution*, 36(7):1359–1372, 2019.
- Fannie I Parra-Cota, Juan J Peña-Cabriales, Sergio de los Santos-Villalobos, Norma A Martínez-Gallardo, and John P Délano-Frier. *Burkholderia ambifaria* and *B. caribensis* promote growth and increase yield in grain amaranth (*Amaranthus cruentus* and *A. hypochondriacus*) by improving plant nitrogen uptake. *PLoS ONE*, 9(2):e88094, 2014.

- Kristian Pastor and M Acanski. The chemistry behind amaranth grains. *Journal of Nutritional Health & Food Engineering*, 8(5):358–360, 2018.
- Sharat Kumar Pradhan, Saumya Ranjan Barik, Ambika Sahoo, Sudipti Mohapatra, Deepak Kumar Nayak, Anumalla Mahender, Jitandriya Meher, Annamalai Anandan, and Elssa Pandit. Population structure, genetic diversity and molecular marker-trait association analysis for high temperature stress tolerance in rice. *PLoS ONE*, 11(8):e0160027, 2016.
- Kim D Pruitt, Tatiana Tatusova, and Donna R Maglott. Ncbi reference sequence (refseq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Research*, 33 (suppl_1):D501–D504, 2005.
- Shaun Purcell, Benjamin Neale, Kathe Todd-Brown, Lori Thomas, Manuel A. R. Ferreira, David Bender, Julian Maller, Pamela Sklar, Paul I. W. de Bakker, Mark J. Daly, and Pak C. Sham. PLINK: A tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics*, 81(3):559–575, 2007. ISSN 0002-9297. doi: 10.1086/519795. URL <https://www.sciencedirect.com/science/article/pii/S0002929707613524>.
- Michael D. Purugganan. Evolutionary insights into the nature of plant domestication. *Current Biology*, 29(14):R705–R714, 2019. ISSN 0960-9822. doi: 10.1016/j.cub.2019.05.053.
- Michael D Purugganan. What is domestication? *Trends in Ecology & Evolution*, 37(8):663–671, 2022.
- Michael D Purugganan and Dorian Q Fuller. The nature of selection during plant domestication. *Nature*, 457(7231):843–848, 2009.
- Mark Ravinet, R Faria, RK Butlin, J Galindo, N Bierne, M Rafajlović, MAF Noor, B Mehlig, and AM Westram. Interpreting the genomic landscape of speciation: a road map for finding barriers to gene flow. *Journal of Evolutionary Biology*, 30(8):1450–1477, 2017.
- Hamid Razifard, Alexis Ramos, Audrey L Valle, Cooper Bodary, Erika Goetz, Elizabeth J Manser, Xiang Li, Lei Zhang, Sofia Visa, Denise Tieman, Esther van der Knaap, and Ana L Caicedo. Genomic evidence for complex domestication history of the cultivated tomato in latin america. *Molecular Biology and Evolution*, 2020. ISSN 0737-4038. doi: 10.1093/molbev/msz297.
- Martha Rendón-Anaya, Josaphat M Montero-Vargas, Soledad Saburido-Álvarez, Anna Vlasova, Salvador Capella-Gutierrez, José Juan Ordaz-Ortiz, O Mario Aguilar, Rosana P Vianello-Brondani, Marta Santalla, Luis Delaye, et al. Genomic history of the origin and domestication of common bean unveils its closest sister species. *Genome Biology*, 18(1):1–17, 2017.

- Jason A Reuter, Damek V Spacek, and Michael P Snyder. High-throughput sequencing technologies. *Molecular Cell*, 58(4):586–597, 2015.
- SS Rich, AE Bell, and SP Wilson. Genetic drift in small populations of tribolium. *Evolution*, pages 579–584, 1979.
- Loren H Rieseberg and John M Burke. The biological reality of species: gene flow, selection, and collective evolution. *Taxon*, 50(1):47–67, 2001.
- Loren H Rieseberg and Jonathan Wendel. Plant speciation: Rise of the poor cousins. *The New Phytologist*, 161(1):3–8, 2004.
- Loren H Rieseberg, C Randal Linder, and Gerald J Seiler. Chromosomal and genic barriers to introgression in *Helianthus*. *Genetics*, 141(3):1163–1171, 1995.
- Kermit Ritland and Michael T Clegg. Evolutionary analysis of plant dna sequences. *The American Naturalist*, 130:S74–S100, 1987.
- Eli Rodgers-Melnick, Peter J Bradbury, Robert J Elshire, Jeffrey C Glaubitz, Charlotte B Acharya, Sharon E Mitchell, Chunhui Li, Yongxiang Li, and Edward S Buckler. Recombination in diverse maize is stable, predictable, and associated with genetic load. *Proceedings of the National Academy of Sciences*, 112(12):3823–3828, 2015.
- Rebekah L Rogers and Montgomery Slatkin. Excess of genomic defects in a woolly mammoth on Wrangel island. *PLoS Genetics*, 13(3):e1006601, 2017.
- Jeffrey Ross-Ibarra, Peter L Morrell, and Brandon S Gaut. Plant domestication, a unique opportunity to identify the genetic basis of adaptation. *Proceedings of the National Academy of Sciences*, 104(suppl 1):8641–8648, 2007.
- Jeffrey Ross-Ibarra, Maud Tenaillon, and Brandon S Gaut. Historical divergence and gene flow in the genus *Zea*. *Genetics*, 181(4):1399–1413, 2009.
- James J Russell, Julie A Theriot, Pranidhi Sood, Wallace F Marshall, Laura F Landweber, Lillian Fritz-Laylin, Jessica K Polka, Snezhana Oliferenko, Therese Gerbich, Amy Gladfelter, et al. Non-model model organisms. *BMC biology*, 15(1):1–31, 2017.
- Jasmine M Saban, Anne J Romero, Thomas HG Ezard, and Mark A Chapman. Extensive crop-wild hybridisation during *Brassica* evolution, and selection during the domestication and diversification of *Brassica* crops. *Genetics*, page iyad027, 2023.

- Fabrice Sagnard, Monique Deu, Dékoro Dembélé, Raphaël Leblois, Lassana Touré, Mohamed Diakité, Caroline Calatayud, Michel Vaxsmann, Sophie Bouchet, Yaya Mallé, et al. Genetic diversity, structure, gene flow and evolutionary relationships within the *Sorghum bicolor* wild–weedy–crop complex in a western African region. *Theoretical and Applied Genetics*, 123:1231–1246, 2011.
- Jonathan D Sauer. Amaranths as dye plants among the pueblo peoples. *Southwestern Journal of Anthropology*, 6(4):412–415, 1950.
- Jonathan D Sauer. The grain amaranths and their relatives: a revised taxonomic and geographic survey. *Annals of the Missouri Botanical Garden*, 54(2):103–137, 1967a.
- Jonathan D. Sauer. The Grain Amaranths and Their Relatives: A Revised Taxonomic and Geographic Survey. *Annals of the Missouri Botanical Garden*, 54(2):103–137, 1967b. ISSN 0026-6493. doi: 10.2307/2394998.
- Jonathan D Sauer. *Historical geography of crop plants: a select roster*. CRC press, 1993.
- Khalid EM Sedeek, Ahmed Mahas, and Magdy Mahfouz. Plant genome engineering for targeted improvement of crop traits. *Frontiers in Plant Science*, 10:114, 2019.
- Eric J Sedivy, Faqiang Wu, and Yoshie Hanzawa. Soybean domestication: the origin, genetic architecture and molecular bases. *New Phytologist*, 214(2):539–553, 2017.
- Andrei Seluanov, Vadim N Gladyshev, Jan Vijg, and Vera Gorbunova. Mechanisms of cancer resistance in long-lived mammals. *Nature Reviews Cancer*, 18(7):433–441, 2018.
- Jason P Sexton, Sharon Y Strauss, and Kevin J Rice. Gene flow increases fitness at the warm edge of a species’ range. *Proceedings of the National Academy of Sciences*, 108(28):11704–11709, 2011.
- B Jesse Shapiro, Jonathan Friedman, Otto X Cordero, Sarah P Preheim, Sonia C Timberlake, Gitta Szabó, Martin F Polz, and Eric J Alm. Population genomics of early events in the ecological differentiation of bacteria. *Science*, 336(6077):48–51, 2012.
- Masoud Sheidai and Fahimeh Koohdar. Evidence for ancient introgression and gene flow in the genus *Tamarix* L. (amaracaceae): a computational approach. *Genetic Resources and Crop Evolution*, pages 1–9, 2023.
- Adam Siepel, Gill Bejerano, Jakob S Pedersen, Angie S Hinrichs, Minmei Hou, Kate Rosenbloom, Hiram Clawson, John Spieth, LaDeana W Hillier, Stephen Richards, et al. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Research*, 15(8):1034–1050, 2005.

- Alexis Simon and Graham Coop. The contribution of gene flow, selection, and genetic drift to five thousand years of human allele frequency change. *bioRxiv*, 2023.
- Mitchell E Skinner, Andrew V Uzilov, Lincoln D Stein, Christopher J Mungall, and Ian H Holmes. JBrowse: a next-generation genome browser. *Genome Research*, 19(9):1630–1638, 2009.
- Montgomery Slatkin. Gene flow in natural populations. *Annual Review of Ecology and Systematics*, 16(1):393–430, 1985.
- Oliver Smith, William V Nicholson, Logan Kistler, Emma Mace, Alan Clapham, Pamela Rose, Chris Stevens, Roselyn Ware, Siva Samavedam, Guy Barker, et al. A domestication history of dynamic adaptation and genomic deterioration in sorghum. *Nature Plants*, 5(4):369–379, 2019.
- Wolfgang Stephan. Selective sweeps. *Genetics*, 211(1):5–13, 2019.
- Markus G Stetter. Limits and constraints to crop domestication. *American Journal of Botany*, 107(12):1617–1621, 2020.
- Markus G Stetter, Leo Zeitler, Adrian Steinhaus, Karoline Kroener, Michelle Biljecki, and Karl J Schmid. Crossing methods and cultivation conditions for rapid production of segregating populations in three grain amaranth species. *Frontiers in Plant Science*, 7:816, 2016.
- Markus G Stetter, Daniel J Gates, Wenbin Mei, and Jeffrey Ross-Ibarra. How to make a domesticate. *Current Biology*, 27(17):R896–R900, 2017a.
- Markus G. Stetter, Thomas Müller, and Karl J. Schmid. Genomic and phenotypic evidence for an incomplete domestication of South American grain amaranth (*Amaranthus caudatus*). *Molecular Ecology*, 26(3):871–886, 2017b. ISSN 1365-294X. doi: 10.1111/mec.13974.
- Markus G Stetter, Mireia Vidal-Villarejo, and Karl J Schmid. Parallel seed color adaptation during multiple domestication attempts of an ancient new world grain. *Molecular Biology and Evolution*, 37(5):1407–1419, 2020.
- Michelle C Stitzer and Jeffrey Ross-Ibarra. Maize domestication and gene interaction. *New Phytologist*, 220(2):395–408, 2018.
- David Swarbreck, Christopher Wilks, Philippe Lamesch, Tanya Z Berardini, Margarita Garcia-Hernandez, Hartmut Foerster, Donghui Li, Tom Meyer, Robert Muller, Larry Ploetz, et al. The arabidopsis information resource (tair): gene structure and function annotation. *Nucleic Acids Research*, 36(suppl_1):D1009–D1014, 2007.

- Fumio Tajima. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics*, 123(3):585–595, 1989.
- A Telschow, J Engelstädter, N Yamamura, P Hammerstein, and GDD Hurst. Asymmetric gene flow and constraints on adaptation caused by sex ratio distorters. *Journal of Evolutionary Biology*, 19(3): 869–878, 2006.
- Maud I Tenaillon, Ewen Burban, Stella Huynh, Arthur Wojcik, Anne-Céline Thuillet, Domenica Manicacci, Pierre R Gérard, Karine Alix, Harry Belcram, Amandine Cornille, et al. Crop domestication as a step toward reproductive isolation. *American Journal of Botany*, 110(7):e16173, 2023.
- Anna Tigano and Vicki L Friesen. Genomics of local adaptation with gene flow. *Molecular Ecology*, 25(10):2144–2164, 2016.
- Joost Van Heerwaarden, John Doebley, William H Briggs, Jeffrey C Glaubitz, Major M Goodman, Jose de Jesus Sanchez Gonzalez, and Jeffrey Ross-Ibarra. Genetic signals of origin, spread, and introgression in a large sample of maize landraces. *Proceedings of the National Academy of Sciences*, 108(3):1088–1092, 2011.
- Rajeev K Varshney, Mahendar Thudi, Manish Roorkiwal, Weiming He, Hari D Upadhyaya, Wei Yang, Prasad Bajaj, Philippe Cubry, Abhishek Rathore, Jianbo Jian, et al. Resequencing of 429 chickpea accessions from 45 countries provides insights into genome diversity, domestication and agronomic traits. *Nature Genetics*, 51(5):857–864, 2019.
- Md Vasimuddin, Sanchit Misra, Heng Li, and Srinivas Aluru. Efficient architecture-aware acceleration of bwa-mem for multicore systems. In *2019 IEEE International Parallel and Distributed Processing Symposium (IPDPS)*, pages 314–324. IEEE, 2019.
- J Craig Venter, Mark D Adams, Eugene W Myers, Peter W Li, Richard J Mural, Granger G Sutton, Hamilton O Smith, Mark Yandell, Cheryl A Evans, Robert A Holt, et al. The sequence of the human genome. *Science*, 291(5507):1304–1351, 2001.
- Jason G Wallace, Eli Rodgers-Melnick, and Edward S Buckler. On the road to breeding 4.0: unraveling the good, the bad, and the boring of crop quantitative genomics. *Annual Review of Genetics*, 2018.
- Baobao Wang, Zechuan Lin, Xin Li, Yongping Zhao, Binbin Zhao, Guangxia Wu, Xiaojing Ma, Hai Wang, Yurong Xie, Quanquan Li, et al. Genome-wide selection and genetic improvement during modern maize breeding. *Nature Genetics*, pages 1–7, 2020a.

- Li Wang, Timothy M Beissinger, Anne Lorant, Claudia Ross-Ibarra, Jeffrey Ross-Ibarra, and Matthew B Hufford. The interplay of demography and selection during maize domestication and expansion. *Genome Biology*, 18(1):1–13, 2017.
- Li Wang, Jiawen Cui, Biao Jin, Jianguo Zhao, Huimin Xu, Zhaogeng Lu, Weixing Li, Xiaoxia Li, Linling Li, Eryuan Liang, et al. Multifeature analyses of vascular cambial cells reveal longevity mechanisms in old ginkgo biloba trees. *Proceedings of the National Academy of Sciences*, 117(4):2201–2210, 2020b.
- Li Wang, Emily B Josephs, Kristin M Lee, Lucas M Roberts, Rubén Rellán-Álvarez, Jeffrey Ross-Ibarra, and Matthew B Hufford. Molecular parallelism underlies convergent highland adaptation of maize landraces. *Molecular Biology and Evolution*, 38(9):3567–3580, 2021.
- GA Watterson. On the number of segregating sites in genetical models without recombination. *Theoretical Population Biology*, 7(2):256–276, 1975.
- Detlef Weigel and Richard Mott. The 1001 genomes project for arabidopsis thaliana. *Genome Biology*, 10(5):1–5, 2009.
- Bryan T Weinstein, Maxim O Lavrentovich, Wolfram Möbius, Andrew W Murray, and David R Nelson. Genetic drift and selection in many-allele range expansions. *PLoS computational biology*, 13(12):e1005866, 2017.
- Bruce S Weir and C Clark Cockerham. Estimating F-statistics for the analysis of population structure. *Evolution*, pages 1358–1370, 1984.
- Howard Wolinsky. The thousand-dollar genome: Genetic brinkmanship or personalized medicine? *EMBO reports*, 8(10):900–903, 2007.
- Sewall Wright. Evolution in mendelian populations. *Genetics*, 16(2):97, 1931.
- Sewall Wright. Genetical structure of populations. *Nature*, 166(4215):247–249, 1950.
- Chung-I Wu and Chau-Ti Ting. Genes and speciation. *Nature Reviews Genetics*, 5(2):114–122, 2004.
- Dongya Wu, Sangting Lao, and Longjiang Fan. De-domestication: an extension of crop evolution. *Trends in Plant Science*, 26(6):560–574, 2021.
- Xiaolei Wu, Chengwei Ren, Trupti Joshi, Tri Vuong, Dong Xu, and Henry T Nguyen. SNP discovery by high-throughput sequencing in soybean. *BMC Genomics*, 11(1):1–10, 2010.

- Yaoyao Wu, Lynn Johnson, Baoxing Song, Cinta Romay, Michelle Stitzer, Adam Siepel, Edward Buckler, and Armin Scheben. A multiple alignment workflow shows the effect of repeat masking and parameter tuning on alignment in plants. *The Plant Genome*, 15(2):e20204, 2022.
- Jian-Long Xu, Jun-Min Wang, Ye-Qing Sun, Li-Jun Wei, Rong-Ting Luo, Ming-Xian Zhang, and Zhi-Kang Li. Heavy genetic load associated with the subspecific differentiation of japonica rice (*Oryza sativa ssp. japonica L.*). *Journal of Experimental Botany*, 57(11):2815–2824, 2006.
- Chin Jian Yang, Luis Fernando Samayoa, Peter J Bradbury, Bode A Olukolu, Wei Xue, Alessandra M York, Michael R Tuholski, Weidong Wang, Lora L Daskalska, Michael A Neumeyer, et al. The genetic architecture of teosinte catalyzed and constrained maize domestication. *Proceedings of the National Academy of Sciences*, 116(12):5643–5652, 2019.
- Ching-chia Yang, Yoshihiro Kawahara, Hiroshi Mizuno, Jianzhong Wu, Takashi Matsumoto, and Takeshi Itoh. Independent domestication of asian rice followed by gene flow from japonica to indica. *Molecular Biology and Evolution*, 29(5):1471–1479, 2012.
- Jian Yang, Teri A Manolio, Louis R Pasquale, Eric Boerwinkle, Neil Caporaso, Julie M Cunningham, Mariza De Andrade, Bjarke Feenstra, Eleanor Feingold, M Geoffrey Hayes, et al. Genome partitioning of genetic variation for complex traits using common snps. *Nature Genetics*, 43(6):519–525, 2011.
- Ning Yang, Yuebin Wang, Xiangguo Liu, Minliang Jin, Miguel Vallebuena-Estrada, Erin Calfee, Lu Chen, Brian P Dilkes, Songtao Gui, Xingming Fan, et al. Two teosintes made modern maize. *Science*, 382(6674):eadg8940, 2023.
- David Zeigler. *Evolution: Components and mechanisms*. Academic Press, 2014.

A Supplementary Information Chapter 2

Supplementary Figures

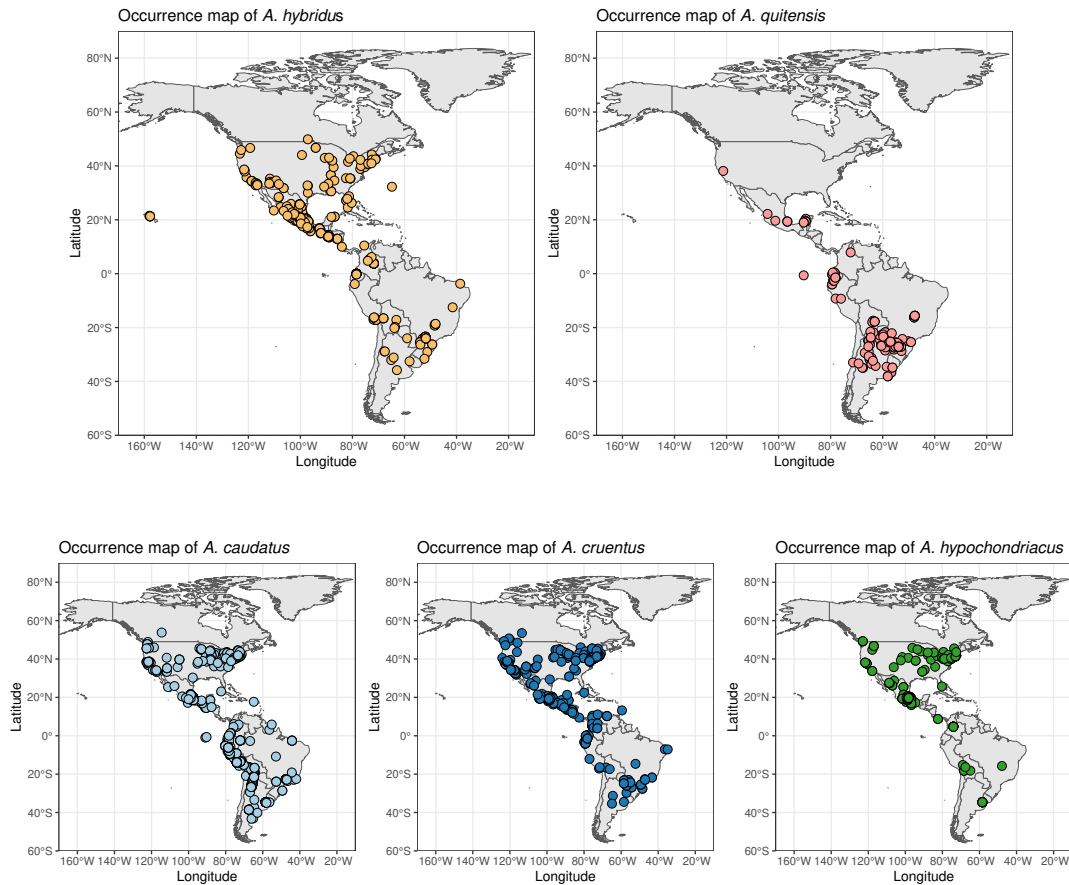


Figure S1: Occurrence map for two wild and three domesticated species of grain amaranth. For *A. hybridus*, the samples in the Central American and South American regions were denoted as *hybridus_CA* and *hybrids_SA*, respectively with no clear range distinction. The data for occurrence were extracted from the GBIF database (<https://doi.org/10.15468/dd.f8cz3g>).

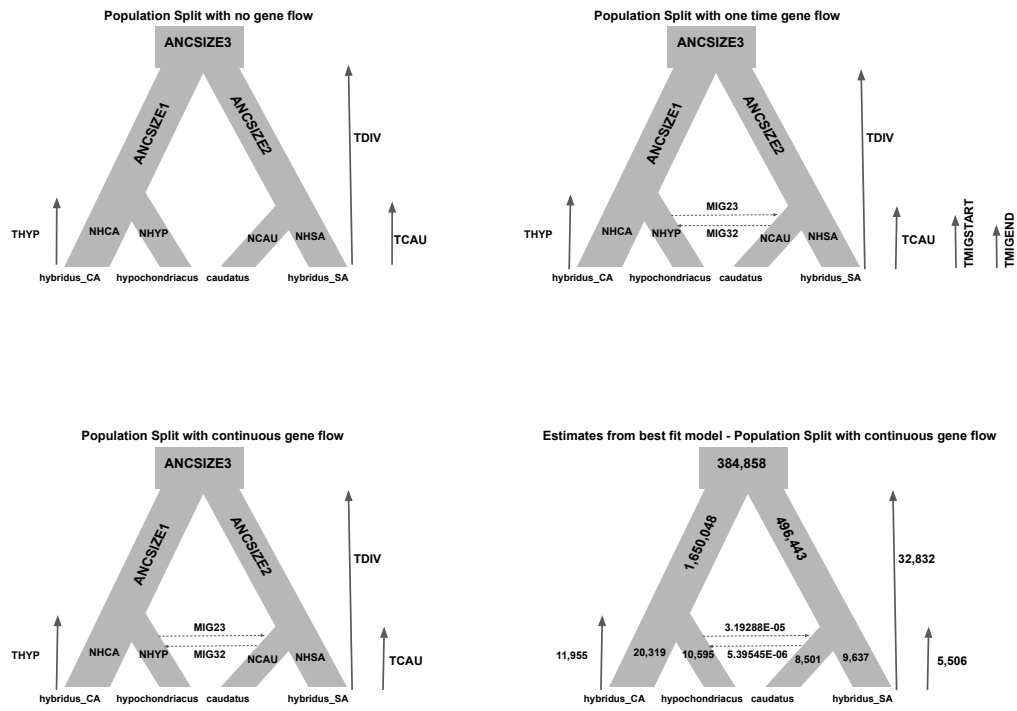


Figure S2: Different demographic models used in Fastsimcoal2 to predict the best scenario. The model of population split with continuous gene flow was predicted to be the best model. Also see Table S2 for detailed demographic parameters for each model.

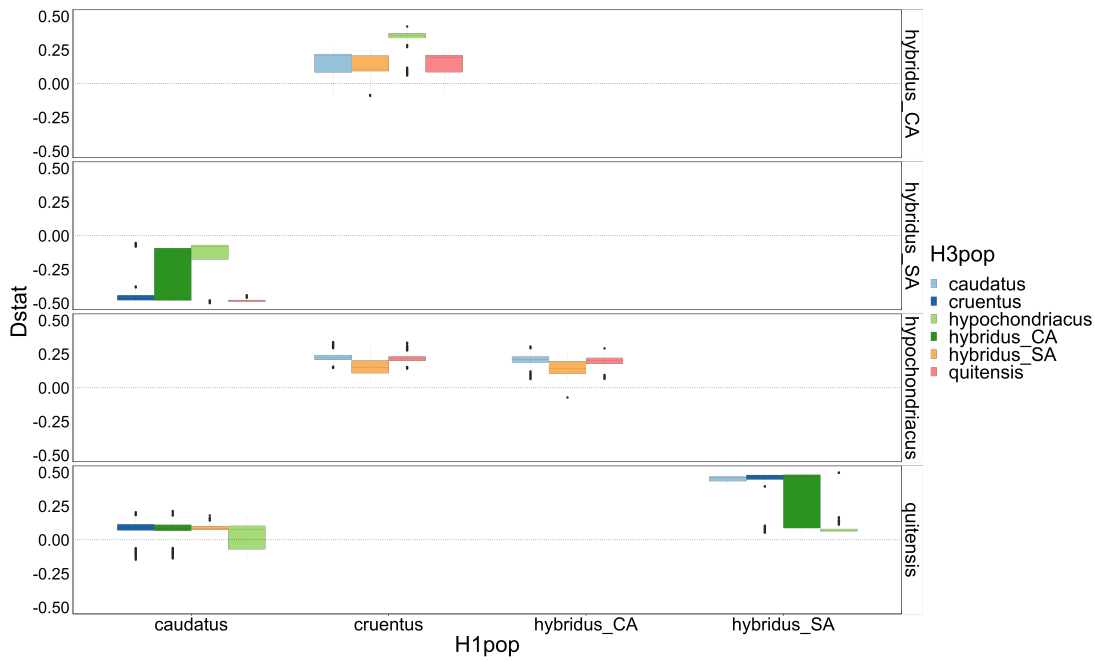


Figure S3: D-statistic value for comparisons between individuals. Each boxplot represents a comparison for a population trios H3 with the inner node H1 and H2. Only significant trees are represented. *A. tuberculatus* was used as an outgroup.

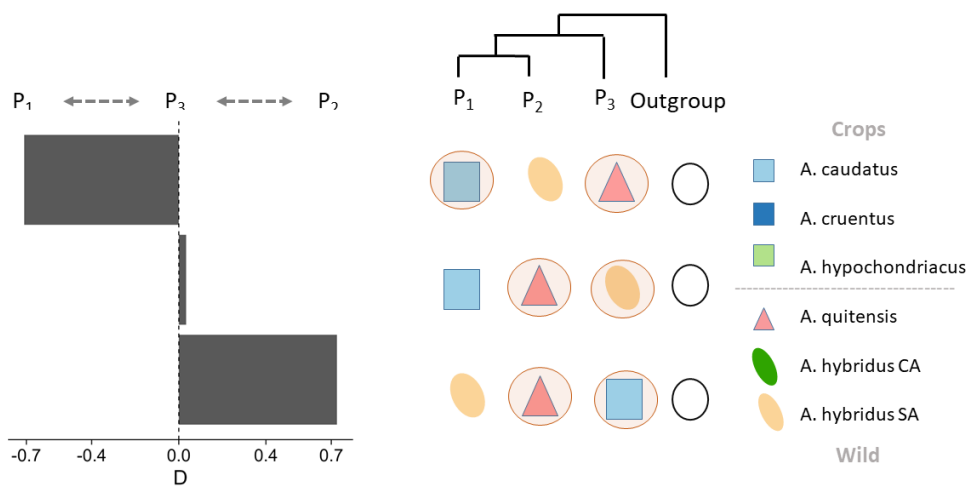


Figure S4: D-statistic value for comparisons between populations of South America. Each bar represents the D-value for each population. The arrows represent the direction of gene flow pairs. Exchanging populations are highlighted with circles. Only significant trees are represented. *A. tuberculatus* was used as an outgroup.

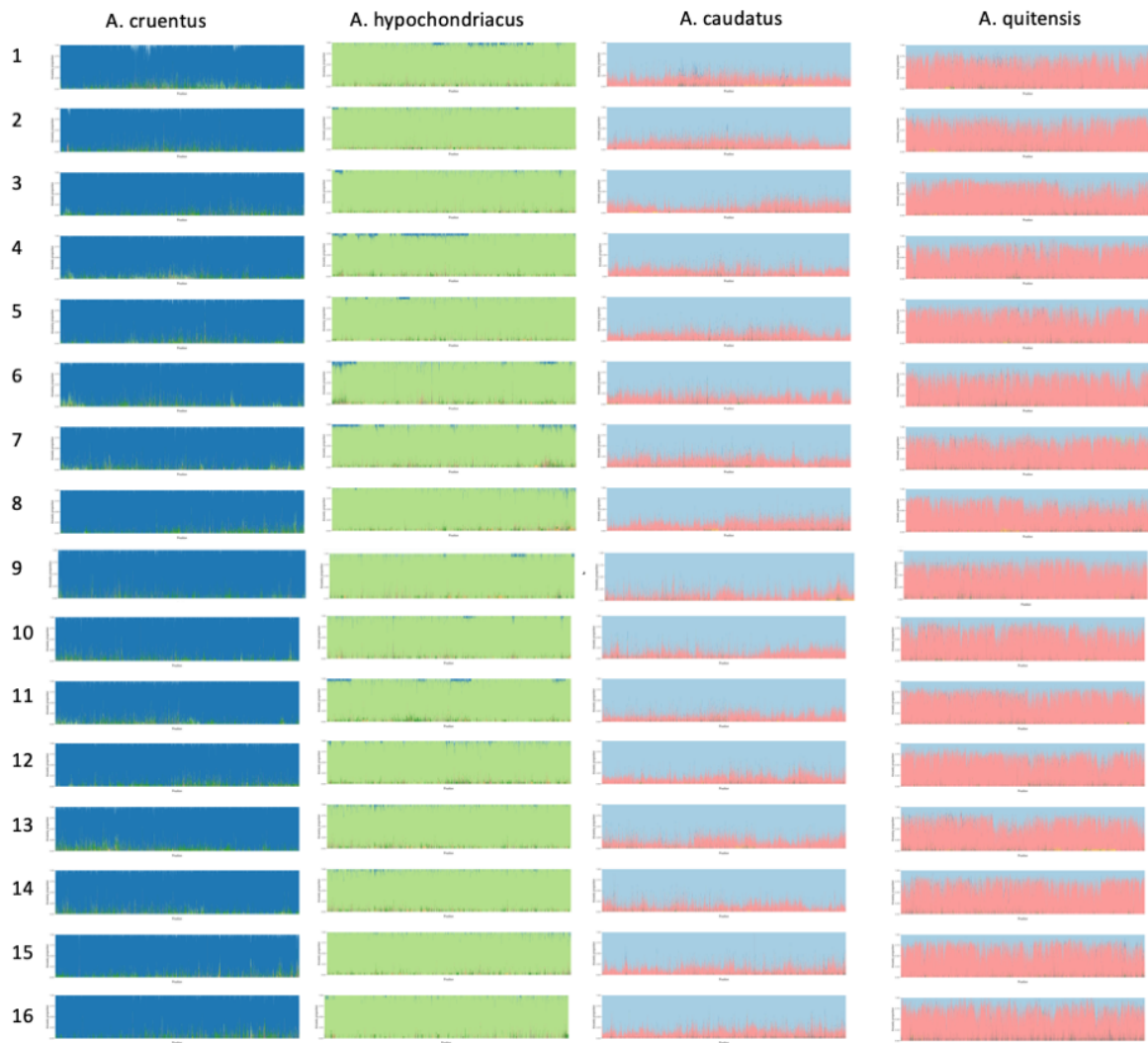


Figure S5: Ancestry proportions summary along the genome, per recipient population, per scaffold. Colors according to Figure 2.

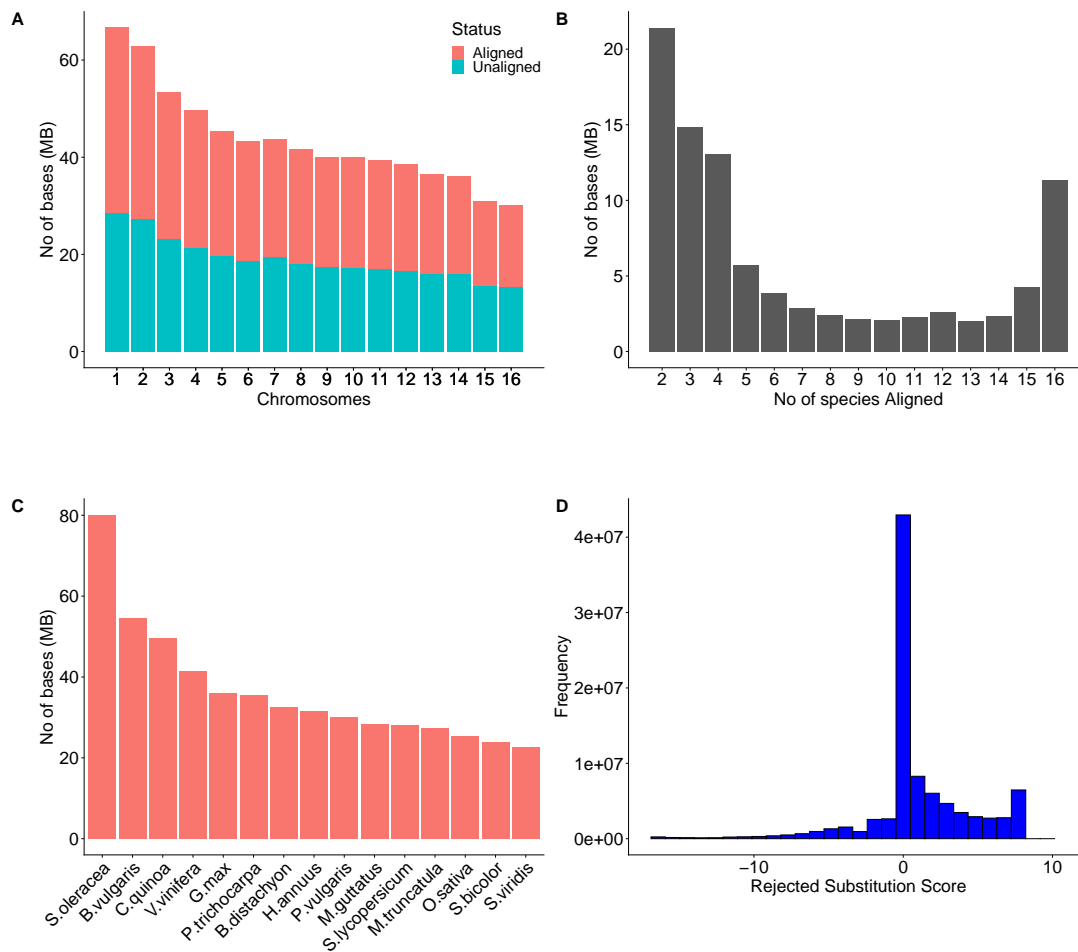


Figure S6: Summary for multiple genome alignment and GERP score. (A) No of bases covered by at least one taxon per Scaffold; (B) Per base coverage of the reference genome by aligning taxon (species); (C) No of bases of the reference genome (*A. hypochondriacus*) covered by each aligning species; (D) Distribution of GERP score (calculated as rejected substitution score).

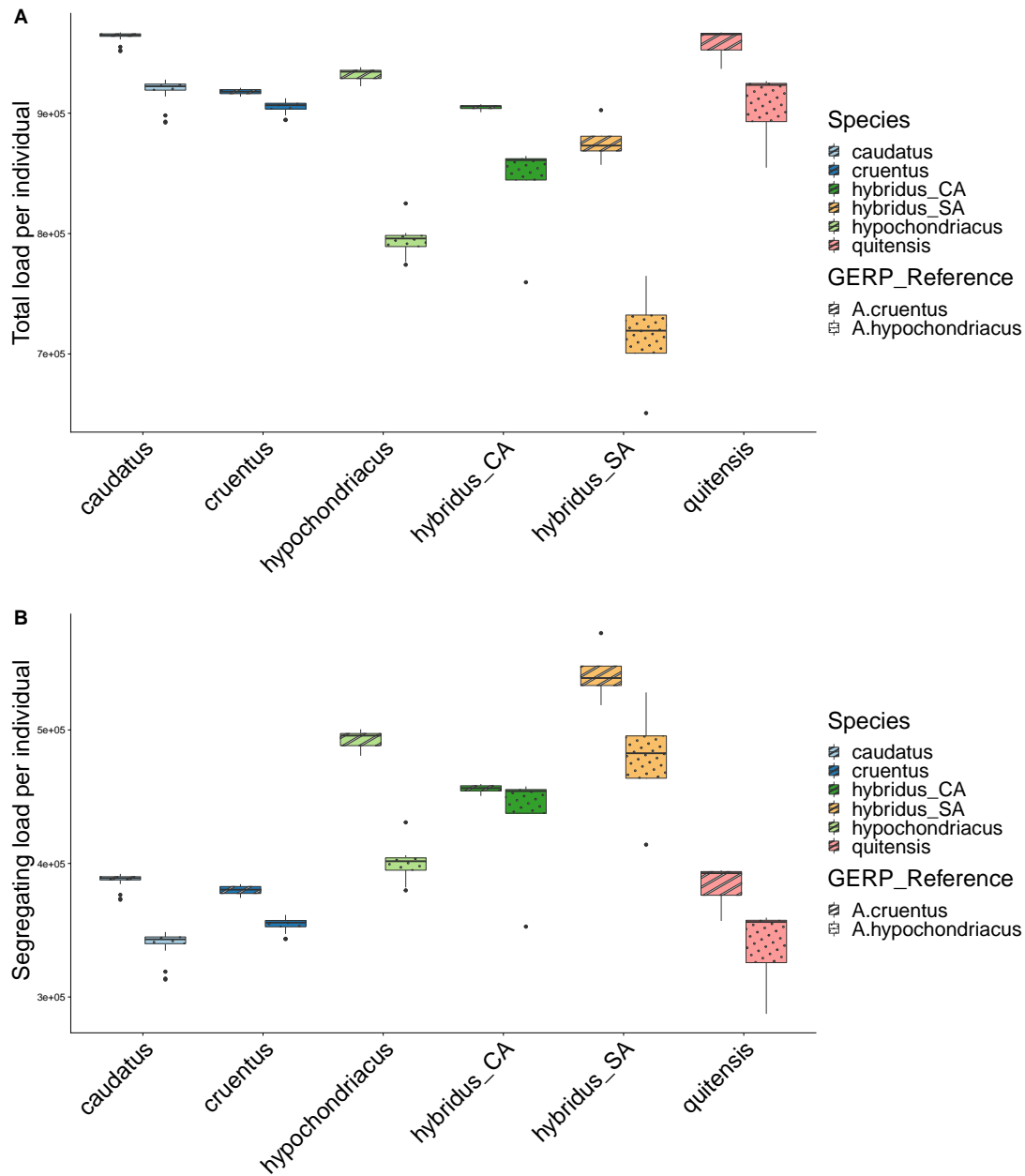


Figure S7: Distribution of genetic load during domestication of grain amaranth using two different genomes for accounting for reference bias. The genetic load was calculated as the sum of GERP scores for derived allele per individual using *A. hypochondriacus* as reference or *A. caudatus* as reference. Distribution of total genetic load per individual in domesticated and wild populations of grain amaranth at (A) all sites (B) segregating sites is depicted using two different reference genomes. The observation of decreased load in hypochondriacus as compared to wild species when using *A. hypochondriacus* genome is not observed when *A. cruentus* is used as a reference for calling GERP. However, the overall pattern for load is comparable for both genomes, suggesting for negligible reference bias.

Supplementary Tables

Table S1: **Summary of non-redundant trees to which we found significant D-values.** Note: ANGSD D-Statistic values were updated to positive, and P1 and P2 positions were switched when negative values were present. This was due to make the comparisons easier with other tools as they require comparisons only between populations on P2 and P3.

P1	P2	P3 (target)	D (ANGSD)	D (Dsuite)	D (admixtools)	f4-ratio	f3 (treemx)
hybridus_SA	quitensis	hypochondriacus	0.225016	NA	NA	NA	NA
caudatus	quitensis	hybridus_SA	0.033597	0.0158596	0.0004726645	0.00997218	0.0887609
caudatus	quitensis	hypochondriacus	0.037009	0.0248318	0.0006999955	0.00776116	0.0831595
caudatus	quitensis	cruentus	0.032232	0.0166697	0.0004753609	0.0037502	0.0755036
cruentus	hybridus_CA	caudatus	0.228626	0.192587	0.0097571493	0.108917	0.0723928
hybridus_CA	hypochondriacus	caudatus	0.10076	0.195984	0.0140320691	0.175783	0.0621802
cruentus	hypochondriacus	caudatus	0.228272	0.286137	0.0237892185	0.265554	0.0578926
hybridus_SA	caudatus	hypochondriacus	0.212658	NA	NA	NA	0.028558
hybridus_SA	quitensis	cruentus	0.17963	NA	NA	NA	0.0255764
hybridus_SA	caudatus	cruentus	0.1691	NA	NA	NA	0.0250975
hybridus_SA	quitensis	caudatus	0.728471	NA	0.0894367462	NA	0.00870881
hybridus_SA	caudatus	quitensis	0.70984	NA	0.0889640817	NA	0.00827429

Table S2: Demographic parameters estimated for different demographic scenarios using Fastsimcoal2.

Parameters	Population Split No gene Flow	Population Split- One time Gene Flow	Population Split Continuous Gene Flow		
			Point Estimates	95 % Confidence Interval	
				Lower Limit	Upper Limit
ANCSIZE1	1571884	1713678	1650048	103347	2408236
ANCSIZE2	1564050	446635	496443	240294	2152774
ANCSIZE3	152842	396708	384858	10335	503711
NHCA	15444	17835	20319	11200	27808
NHSA	16529	16548	9637	8736	22814
NHYP	14762	8349	10595	5838	13133
NCAU	15146	12949	8501	7099	14691
THYP	8728	9894	11955	5916	16802
TCAU	9553	9633	5506	5065	12818
TDIV	31036	30952	32832	30405	40153
TMIGSTART	-	9157	-	-	-
TMIGEND	-	9344	-	-	-
MIG23	-	3.38801E-05	3.19288E-05	2.4E-05	4.7E-05
MIG32	-	3.45086E-06	5.39545E-06	2E-06	9E-06
MaxEstLhood	-83545622.13	-82852846.15	-82822883.13	-	-
deltaL	8294205.304	7601429.324	7571466.3	-	-
AIC	384741828.2	381551484.9	381413496.1	-	-
deltaAIC	3328332.113	137988.8248	0	-	-

ANCSIZE1 represents ancestral population size of hypochondriacus and hybridus_CA.

ANCSIZE2 represents ancestral population size of caudatus and hybridus_SA.

ANCSIZE3 represents Ancestral population size.

NHCA represents the population size of hybridus_CA.

NHSA represents the population size of hybridus_SA.

NHYP represents the population size of hypochondriacus.

NCAU represents the population size of caudatus.

THYP represents time of split for hypochondriacus

TCAU represents time of split for caudatus

TDIV represents divergence time between hybridus_CA and hybridus_SA

TMIGSTART represents the start time of the migration

TMIGEND represents the end time of the first migration

MIG23 represents migration rate from hypochondriacus to caudatus

MIG32 represents migration rate from caudatus to hypochondriacus

MaxEstLhood maximum log-likelihood of the best estimate.

deltaL difference between estimated and observed log-likelihood.

AIC Akaike's information criterion, $AIC = 2d - 2\ln(Lhood)$, where d is the number of parameters.

deltaAIC $AIC - \min(AIC)$.

Table S3: Crosses within and between species along with primer pairs used to validate the crosses. The "intra" represents crosses of plants from the same species (different accessions); "inter" represents crosses between plants of different species.

crossing type	maternal species	paternal species	maternal accession	paternal accession	survival rate (%)	primer pair
inter	A. caudatus	A. cruentus	PI 490518	PI 511714	0	Primer5
inter	A. caudatus	A. cruentus	PI 490518	PI 643058	0	Primer5
inter	A. caudatus	A. cruentus	PI 490612	PI 511714	0	Primer5
inter	A. caudatus	A. cruentus	PI 490518	PI 643058	0	Primer5
inter	A. caudatus	A. cruentus	PI 490518	PI 511717	0	Primer5
inter	A. caudatus	A. cruentus	PI 490612	PI 643058	0	Primer6
inter	A. caudatus	A. hypochondriacus	PI 490518	PI 558499	100	Primer5
inter	A. caudatus	A. hypochondriacus	PI 490518	PI 643070	100	Primer5
inter	A. caudatus	A. hypochondriacus	PI 490612	PI 643070	100	Primer6
inter	A. caudatus	A. hypochondriacus	PI 490612	PI 558499	100	Primer6
inter	A. caudatus	A. hypochondriacus	PI 490518	PI 558499	100	Primer5
inter	A. hypochondriacus	A. cruentus	PI 643070	PI 511714	100	Primer3
inter	A. hypochondriacus	A. cruentus	PI 558499	PI 643058	100	Primer3
inter	A. hypochondriacus	A. cruentus	PI 643070	PI 643058	0	Primer3
intra	A. caudatus	A. caudatus	PI 642741	PI 490518	100	Primer5
intra	A. caudatus	A. caudatus	PI 490612	PI 490518	100	Primer5
intra	A. caudatus	A. caudatus	PI 490612	PI 642741	100	Primer6
intra	A. cruentus	A. cruentus	PI 643058	PI 511714	100	Primer3
intra	A. cruentus	A. cruentus	PI 511714	PI 643058	100	Primer4
intra	A. cruentus	A. cruentus	PI 511717	PI 511714	100	Primer2
intra	A. cruentus	A. cruentus	PI 511717	PI 643058	100	Primer4
intra	A. hypochondriacus	A. hypochondriacus	PI 558499	PI 604581	100	Primer1
intra	A. hypochondriacus	A. hypochondriacus	PI 604581	PI 558499	100	Primer1
intra	A. hypochondriacus	A. hypochondriacus	PI 643070	PI 558499	100	Primer3
intra	A. hypochondriacus	A. hypochondriacus	PI 604581	PI 643070	100	Primer1

Table S4: List of primers used to validate the crosses for the F1 plants

primer pair	forward primer	reverse primer
Primer1	TCACCAATCCCTCCCTCCAA	ACGCGGCGGTTATATGTGAT
Primer2	ACAATTCACATGCAAGCCGG	CCCGTTGCACGATTTTCCAA
Primer3	GACTTGCCTCCTGGAATGCA	AAATCGGTGCAACGTTCTGC
Primer4	GTGACGACAATGATGCTGCC	CGTAACGCATGTGGCATCTG
Primer5	AGTAGACAAACTGGAACCCGA	TGGTCACTTCCAAGGTATGCA
Primer6	AGCTTGTTCAATGCATGGGT	ACGCAACTCTTACAGAGGTCG

Table S5: List of accessions used in this the study. Names follow USDA germplasm ID

Name	Country_Origin	Population	ENA_ID
PI 490689	Ecuador	caudatus	ERR3021332
PI 490739	Ecuador	caudatus	ERR3021337
PI 490511	Peru	caudatus	ERR3021351
PI 490491	Argentina	caudatus	ERR3021382
PI 481949	Peru	caudatus	ERR3021403
PI 490459	Bolivia	caudatus	ERR3021366
PI 481957	Peru	caudatus	ERR3021380
PI 511706	Peru	caudatus	ERR3021381
AMA 125	Peru	caudatus	ERR3021385
PI 490561	Peru	caudatus	ERR3021387
PI 642741	Bolivia	caudatus	ERR3021388
PI 511704	Peru	caudatus	ERR3021389
PI 490518	Peru	caudatus	ERR3021390
PI 649227	Peru	caudatus	ERR3021394
PI 481960	Peru	caudatus	ERR3021395
PI 511687	Peru	caudatus	ERR3021397
PI 490612	Peru	caudatus	ERR3021398
PI 649217	Peru	caudatus	ERR3021399
PI 511712	Peru	caudatus	ERR3021400
PI 490431	Peru	caudatus	ERR3021401
PI 490609	Ecuador	caudatus	ERR3021402
PI 511696	Peru	caudatus	ERR3021404
PI 511686	Peru	caudatus	ERR3021405
PI 481965	Peru	caudatus	ERR3021406
PI 511690	Peru	caudatus	ERR3021407
PI 686455	Peru	caudatus	ERR3021408
PI 490604	Bolivia	caudatus	ERR3021409
PI 511679	Argentina	caudatus	ERR3021430
PI 511680	Argentina	caudatus	ERR3021431
PI 511681	Bolivia	caudatus	ERR3021432

PI 608019	Ecuador	caudatus	ERR3021440
PI 649228	Peru	caudatus	ERR3021443
PI 649230	Peru	caudatus	ERR3021444
PI 667165	Brazil	cruentus	ERR3021328
PI 649509	Mexico	cruentus	ERR3021329
PI 643042	Mexico	cruentus	ERR3021422
PI 643039	Mexico	cruentus	ERR3021410
PI 643058	Mexico	cruentus	ERR3021411
PI 511713	Peru	cruentus	ERR3021412
PI 511717	Guatemala	cruentus	ERR3021413
PI 649514	Mexico	cruentus	ERR3021414
PI 606798	Mexico	cruentus	ERR3021415
PI 576482	Mexico	cruentus	ERR3021416
PI 649609	Mexico	cruentus	ERR3021417
PI 451826	Guatemala	cruentus	ERR3021419
Ames5552	Mexico	cruentus	ERR3021421
PI 643049	Mexico	cruentus	ERR3021423
PI 649524	Mexico	cruentus	ERR3021424
PI 433228	Guatemala	cruentus	ERR3021429
PI 511714	Peru	cruentus	ERR3021433
PI 576481	Mexico	cruentus	ERR3021436
PI 643037	Mexico	cruentus	ERR3021442
PI 658728	Mexico	cruentus	ERR3021447
PI 667160	Guatemala	cruentus	ERR3021448
PI 667158	Guatemala	hybridus_CA	ERR3021331
PI 511724	Mexico	hybridus_CA	ERR3021336
PI 604582	Mexico	hybridus_CA	ERR3021340
PI 604574	Mexico	hybridus_CA	ERR3021341
PI 604568	Mexico	hybridus_CA	ERR3021437
PI 490489	Peru	hybridus_SA	ERR3021333
PI 511754	Ecuador	hybridus_SA	ERR3021391
PI 686451	Peru	hybridus_SA	ERR3021426
PI 636180	Colombia	hybridus_SA	ERR3021441
PI 649537	Mexico	hypochondriacus	ERR3021343

PI 643036	Mexico	hypochondriacus	ERR3021344
PI 649575	Mexico	hypochondriacus	ERR3021345
Ames5457	Mexico	hypochondriacus	ERR3021346
PI 633589	Mexico	hypochondriacus	ERR3021347
PI 649565	Mexico	hypochondriacus	ERR3021348
PI 604559	Mexico	hypochondriacus	ERR3021349
PI 643070	Mexico	hypochondriacus	ERR3021350
PI 604581	Mexico	hypochondriacus	ERR3021352
Ames2085	Mexico	hypochondriacus	ERR3021353
PI 643041	Mexico	hypochondriacus	ERR3021354
PI 649602	Mexico	hypochondriacus	ERR3021355
PI 604587	Mexico	hypochondriacus	ERR3021356
PI 649607	Mexico	hypochondriacus	ERR3021357
PI 511731	Mexico	hypochondriacus	ERR3021359
PI 643067	Mexico	hypochondriacus	ERR3021360
PI 649559	Mexico	hypochondriacus	ERR3021361
PI 649595	Mexico	hypochondriacus	ERR3021362
PI 649623	Mexico	hypochondriacus	ERR3021364
PI 604595	Mexico	hypochondriacus	ERR3021439
PI 649529	Mexico	hypochondriacus	ERR3021445
PI 511749	Ecuador	quitensis	ERR3021365
PI 511737	Ecuador	quitensis	ERR3021367
PI 490664	Ecuador	quitensis	ERR3021335
PI 667156	Ecuador	quitensis	ERR3021338
PI 490705	Ecuador	quitensis	ERR3021370
PI 490684	Ecuador	quitensis	ERR3021342
PI 691596	Argentina	quitensis	ERR3021368
PI 649246	Peru	quitensis	ERR3021369
PI 511745	Ecuador	quitensis	ERR3021374
PI 652426	Brazil	quitensis	ERR3021371
PI 490720	Ecuador	quitensis	ERR3021376
PI 511741	Ecuador	quitensis	ERR3021377
PI 669830	Peru	quitensis	ERR3021378
PI 511751	Peru	quitensis	ERR3021372

PI 652428	Brazil	quitensis	ERR3021373
PI 652422	Brazil	quitensis	ERR3021375
PI 490466	Peru	quitensis	ERR3021392
PI 490673	Ecuador	quitensis	ERR3021393
PI 669836	Argentina	quitensis	ERR3021379
PI 490670	Ecuador	quitensis	ERR3021334
PI 490679	Ecuador	quitensis	ERR3021396
PI 669839	Peru	quitensis	ERR3021428
PI 511736	Bolivia	quitensis	ERR3021434
PI 511747	Ecuador	quitensis	ERR3021435
ERR3220318		tuberculatus	ERR3220318

B Supplementary Information Chapter 3

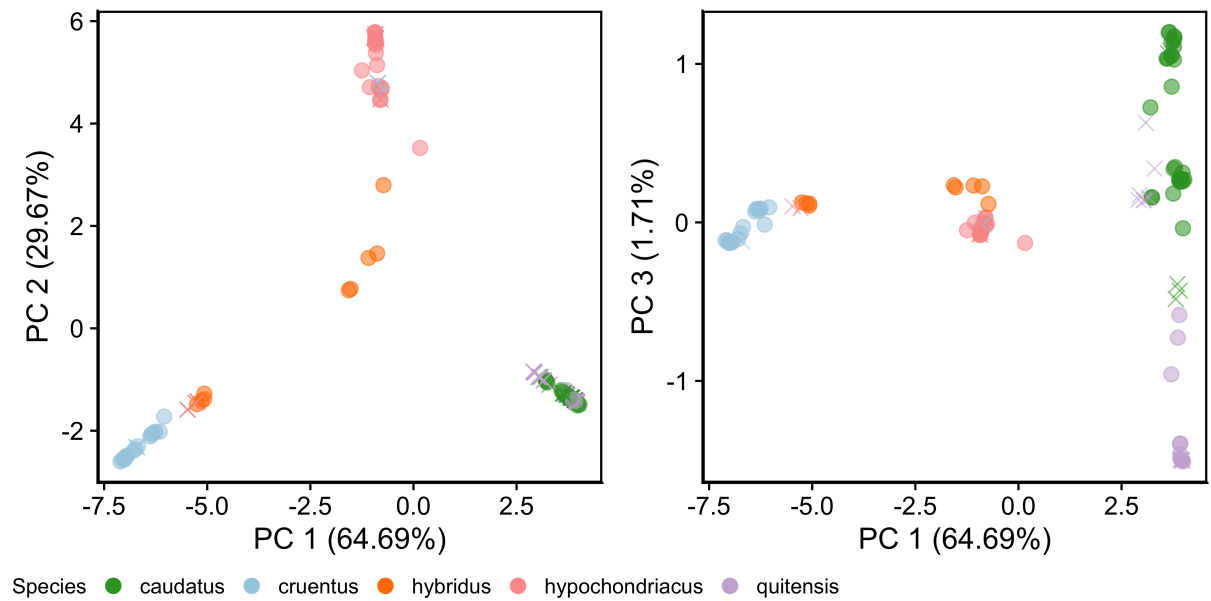


Figure S1: Each symbol represents each of the 116 sample. Circles represent amaranth samples included in the study. Removed samples are marked with crosses. *A. caudatus* (green), *A. cruentus* (blue), *A. hybridus* (orange), *A. hypochondriacus* (rose), *A. quitensis* (purple). Axis show the percentage of variance explained by each principal component.

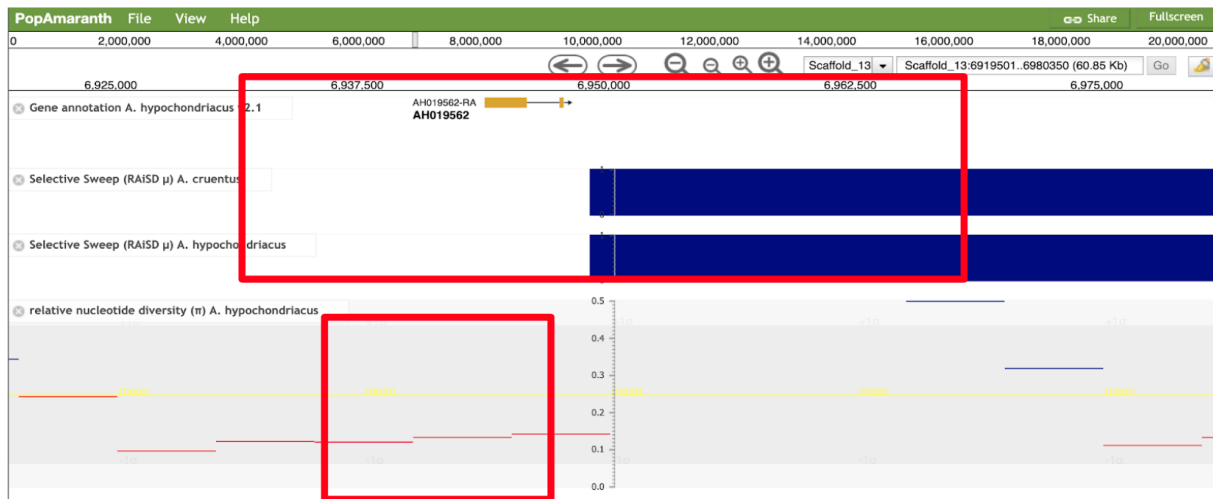


Figure S2: .

Signals of positive selection identified in *A. cruentus* and *A. hypochondriacus*. Relative nucleotide diversity between *A. hypochondriacus* compared to its wild ancestor *A. hybridus* is lower than the genome wide relative diversity, which is an indicator of selection in this region.

Table S1: List of samples evaluated. Samples marked with * were filtered and not included in PopAmaranth

Accession name	Taxonomy	Origin
PI576481	cruentus	Mexico
PI490670 *	quitensis	Ecuador
PI649623	hypochondriacus	Mexico
PI481960	caudatus	Peru
PI511751 *	quitensis	Peru
PI490705	quitensis	Ecuador
AMA155	cruentus	USA
PI490491 *	caudatus	Argentina
PI511717	cruentus	Guatemala
PI433228	cruentus	Guatemala
AMA125	caudatus	Peru
PI511680	caudatus	Argentina
PI649230	caudatus	Peru
PI481949 *	caudatus	Peru
Ames5247	quitensis	Peru
PI511724	hybridus	Mexico
PI642741	caudatus	Bolivia
Ames5232	hybridus	Peru
PI608019	caudatus	Ecuador
PI649565 *	hypochondriacus	Mexico
PI649217	caudatus	Peru
PI511706	caudatus	Peru
PI649607	hypochondriacus	Mexico

Table S1 – *Continued from previous page*

Accession name	Taxonomy	Origin
PI490664 *	quitensis	Ecuador
PI633589	hypochondriacus	Mexico
PI511737	quitensis	Ecuador
PI511690	caudatus	Peru
PI511687	caudatus	Peru
PI511679	caudatus	Argentina
PI649537	hypochondriacus	Mexico
PI511696	caudatus	Peru
PI490466	quitensis	Peru
PI490459	caudatus	Bolivia
PI604587	hypochondriacus	Mexico
PI643042 *	cruentus	Mexico
PI511704	caudatus	Peru
PI511713	cruentus	Peru
PI652432 *	hypochondriacus	Brazil
PI649602	hypochondriacus	Mexico
PI511745	quitensis	Ecuador
PI643039	cruentus	Mexico
PI511749	quitensis	Ecuador
PI490679 *	quitensis	Ecuador
PI511686	caudatus	Peru
PI481957	caudatus	Peru
PI511741	quitensis	Ecuador
PI667156 *	quitensis	Ecuador
PI490739 *	caudatus	Ecuador

Table S1 – *Continued from previous page*

Accession name	Taxonomy	Origin
PI490720	quitensis	Ecuador
PI490731 *	quitensis	Ecuador
PI604581	hypochondriacus	Mexico
PI649514	cruentus	Mexico
PI649587 *	hypochondriacus	Mexico
PI511723 *	cruentus	Mexico
PI649559	hypochondriacus	Mexico
PI643067	hypochondriacus	Mexico
PI636180	hybridus	Colombia
PI643049	cruentus	Mexico
PI511681	caudatus	Bolivia
PI490609	caudatus	Ecuador
PI649246 *	quitensis	Peru
PI604568	hybridus	Mexico
PI658728	cruentus	Mexico
PI511731	hypochondriacus	Mexico
Ames21666 *	quitensis	Argentina
Ames5457	hypochondriacus	Mexico
PI643070	hypochondriacus	Mexico
PI490673	quitensis	Ecuador
PI643037	cruentus	Mexico
PI643036	hypochondriacus	Mexico
Ames5334 *	quitensis	Argentina
Baernkrafft *	hypochondriacus	Germany
PI490511 *	caudatus	Peru

Table S1 – *Continued from previous page*

Accession name	Taxonomy	Origin
PI643041	hypochondriacus	Mexico
PI490684 *	quitensis	Ecuador
PI649228	caudatus	Peru
PI649529	hypochondriacus	Mexico
PI490518	caudatus	Peru
PI604582	hybridus	Mexico
PI652422 *	quitensis	Brazil
PI604574	hybridus	Mexico
PI490431	caudatus	Peru
PI649595 *	hypochondriacus	Mexico
PI649609	cruentus	Mexico
Ames5302	caudatus	Peru
Ames5342	quitensis	Peru
PI511876 *	cruentus	Mexico
PI649509	cruentus	Mexico
PI490561	caudatus	Peru
PI490612	caudatus	Peru
PI649227	caudatus	Peru
PI667158	hybridus	Guatemala
PI606798	cruentus	Mexico
Ames2085	hypochondriacus	Mexico
PI511747	quitensis	Ecuador
PI604559 *	hypochondriacus	Mexico
PI511754	hybridus	Ecuador
Ames5552	cruentus	Mexico

Table S1 – *Continued from previous page*

Accession name	Taxonomy	Origin
PI490604	caudatus	Bolivia
PI667165	cruentus	Brazil
PI481965	caudatus	Peru
PI511714	cruentus	Peru
PI451826	cruentus	Guatemala
PI667160	cruentus	Guatemala
PI652428 *	quitensis	Brazil
PI511736	quitensis	Bolivia
PI490689 *	caudatus	Ecuador
Ames2215 *	hypochondriacus	Mexico
PI649575	hypochondriacus	Mexico
PI652426 *	quitensis	Brazil
PI576482	cruentus	Mexico
PI643058	cruentus	Mexico
PI490489	hybridus	Peru
PI649524	cruentus	Mexico
PI604595	hypochondriacus	Mexico
PI511712	caudatus	Ecuador

Continued on next page

Table S2: **List of all tracks available on PopAmaranth at the time of publication.** Detailed description of the included categories (bold) and respective tracks and summary statistics.

Track Name	Description	Detailed explanation
Annotation		
Reference Genome v2.0	<i>Amaranthus hypochondriacus</i> reference genome v2.0 (Lightfoot et al., 2017a)	Reference sequence of the <i>A. hypochondriacus</i> (version 2.0 from Phytozome) (Goodstein et al., 2012)
Gene Annotation v2.1	<i>Amaranthus hypochondriacus</i> gene annotation with subfeatures.	Follows Phytozome nomenclature for gene names. Clicking on the track, information for subfeatures CDS, mRNA, and UTR's is present. For each of the subfeatures, its name, type, position on the chromosome, and length are described. <i>Gene density</i> viewable from whole chromosome perspective.
Differentiation		
F_{st}	Weir-Cockerham pairwise F_{st} (Weir and Cockerham, 1984)	Pairwise F_{st} between species pairs calculated on non-overlapping 5 kb windows. The yellow line represent the global mean F_{st} . Windows with F_{st} below the mean are represented in red and above in blue. Shading in light and darker grey represents 1 standard deviation and 2 standard deviations from the mean, respectively. Maximum and minimum scale is adjusted for the local region in display.

Continued on next page

Table S2 – *Continued from previous page*

Track Name	Description	Detailed explanation
		<p><i>Summary F_{st}</i>: Color gradient showing F_{st} values for all comparisons. Color scale varies from white (0) to dark blue (1). Each segment represents the value of a 5kb window.</p>
Diversity		
Wu & Watterson θ	Estimator of genetic diversity (Watterson, 1975)	<p>θ for each pairs calculated in 5 kb non-overlapping windows. The yellow line represent the global mean θ. Windows below the genome-wide mean are represented in red and above the mean in blue. Shading in light and dark grey represents 1 standard deviation and 2 standard deviations from the mean, respectively. Lower θ indicates lower genetic diversity within the population. Maximum and minimum scale is adjusted for the local region in display.</p> <p><i>Summary Wu & Watterson θ</i>: Color gradient showing θ values per species. Color scale varies from white (0) to dark blue (maximum). Each segment represents the value of a 5kb window.</p>

Continued on next page

Table S2 – *Continued from previous page*

Track Name	Description	Detailed explanation
Expected heterozygosity	Expected heterozygosity for a SNP under Hardy-Weinberg equilibrium	<p>The yellow line represent the global mean expected heterozygosity. SNP based windows are shown in blue when above the mean and in red when below the mean. Values range from 0 to 0.5. Maximum and minimum scale is adjusted for the local region in display.</p> <p><i>Summary expected heterozygosity:</i> Color gradient showing expected heterozygosity values per species. Color scale varies from white (0) to dark blue in the maximum vale (0.5). Each segment represents the value for a variant.</p>
Observed heterozygosity	Observed heterozygosity for a SNP genotype.	<p>The yellow line represents the genome-wide mean observed heterozygosity. SNP-based observed heterozygosity are represented in blue when above the mean and in red when below the mean. Each point represents the value for one variant. Values range from 0 to 1. Maximum and minimum scale is adjusted for the local region in display.</p> <p><i>Summary observed heterozygosity:</i> Color gradient showing observed heterozygosity values per species. Color scale varies from white (0) to dark blue in the maximum vale (1). Each segment represents the value for one variant.</p>

Continued on next page

Table S2 – *Continued from previous page*

Track Name	Description	Detailed explanation
Inbreeding coefficient	Inbreeding coefficient for each variant.	<p>Calculated based on observed heterozygosity and expected heterozygosity. The yellow line represent the global mean inbreeding coefficient. SNP-based inbreeding coefficients are represented in blue when above the mean and in red when below the mean. Values range from dark red (minimum, -1) to white (0) to dark blue (maximum, 1). Maximum and minimum scale is adjusted for the local region in display.</p> <p><i>Summary inbreeding coefficient:</i> Color gradient showing inbreeding coefficient values. Color scale varies from in the minimum vale white (-1) to darker blue in the maximum value (1). Each segment represents the value for one variant.</p>
Nucleotide diversity (π)	Nucleotide diversity (Nei and Li, 1979)	<p>Average nucleotide diversity (π) in 5kb windows. The yellow line represent the genome-wide mean π. Values above the mean are represented in blue, values below the genome-wide mean are red. Shading in light and darker grey represents 1 standard deviation and 2 standard deviations from the mean, respectively. Maximum and minimum scale is adjusted for the local region in display.</p>

Continued on next page

Table S2 – *Continued from previous page*

Track Name	Description	Detailed explanation
		<i>Summary nucleotide diversity (π):</i> Color gradient showing π values for each species. Color scale varies from white (0) to darker blue (maximum). Each segment represents the value of a 5kb window.
Selection		
Tajima's D	Scaled difference between the mean number of pairwise differences and the number of segregating sites Tajima (1989)	Deviations from neutral state ($D=0$) are possible signals of selection or demographic changes. Tajima's D values for windows of 5kb. The yellow line represent the genome-wide mean π . Values above the mean are represented in blue, values below the genome-wide mean are red. Shading in light and darker grey represents 1 standard deviation and 2 standard deviations from the mean, respectively. Maximum and minimum scale is adjusted for the local region in display. <i>Summary Tajima's D:</i> Color gradient showing Tajima's D values for each species. Color scale varies from dark red (minimum, <0) to white (0) to dark blue (maximum, >0). Each segment represents the value of one 5kb window.

Continued on next page

Table S2 – *Continued from previous page*

Track Name	Description	Detailed explanation
Relative nucleotide diversity	Ratio of nucleotide diversity of a domesticated species and their wild ancestor (<i>A. hybridus</i>)	Tracks are available for each of the three grain amaranth species (<i>A. caudatus</i> , <i>A. cruentus</i> , <i>A. hypochondriacus</i>) relative to the wild ancestor (<i>A. hybridus</i>) in 5 kb windows. The yellow line denotes the genome wide average. Windows above the mean are represented in blue and windows below the mean in red. Maximum and minimum scale is adjusted for the local region in display. <i>Summary relative nucleotide diversity:</i> Color gradient showing relative π values for each species. Color scale varies from dark red (minimum, <1) to white (1) to dark blue (maximum, >1)
Selective Sweeps (RAiSD (μ))	μ statistic for selective sweep detection	Each blue window represents the extend of a selective sweep, given by merged windows of 20 SNPs within the top 1% of μ values. These outliers indicate regions with signals of positive selection. <i>Summary Selective Sweeps (RAiSD (μ)):</i> Segments in blue show selective sweeps for all five species.
Variant Call		

Continued on next page

Table S2 – *Continued from previous page*

Track Name	Description	Detailed explanation
VCF	Single nucleotide polymorphisms (SNP) identified.	Variant calls from Stetter et al. (2020) within each amaranth species. Genotypes are displayed in pie charts. Gold colored slices represent the percentage of homozygous reference genotypes, green all other genotypes (heterozygous and homozygous alternative).

Acknowledgments

First and foremost, I would like to thank my supervisor, Prof. Markus G. Stetter. It was a special experience to see the growth of the group alongside you. I might or might not have contributed to some of those new grey hairs that appeared over the years, but I want to take this opportunity to thank you for the constant support that you gave me. There were so many circumstances that we always managed to adapt to, and if today I am a better scientist than I was at the beginning of this journey, that is mostly thanks to you. There are so many invaluable, small, foundational things that I see myself repeating every day now that I got from you. I would like to thank Prof. Thomas Wiehe and Prof. Juliette de Meaux for participating in my thesis committee and providing guidance over the years.

My co-authors, Corbinian Graf and Dr. Akanshka Singh, for their great contributions to the new research questions we could answer together. I also want to thank the PPIG for fostering my interest in Population Genetics and creating opportunities for discussion and knowledge sharing. A word for the Hülskamp group as well, for the seminars that open discussions on botany context and the daily collaborative spirit. Also, I would like to thank the de Meaux group for the weekly discussions that helped push the research further. A special word to Margarita for the constant willingness to help coupled with good-humour complaining. I want to thank the remaining team on the Stetter Lab, who were always willing to help each other and had a welcoming spirit, especially Tom, whom I followed from his Bachelor's until becoming a PhD student himself. I wish you all the success.

Uma palavra muito especial à minha família. O caminho nem sempre foi uma linha recta, mas estiveram sempre lá para me apoiar em todos os momentos, fossem eles mais ou menos complicados. As fundações e os princípios começam convosco e se hoje estou nesta posição, devo-o muito a vocês. Obrigado!

And last, but most importantly, to Sara. There are not enough words to describe your unwavering support, combined with such stubbornness and kind heart. It is such a joy to share this roller-coaster of experiences with you, and I cannot feel more lucky to have such a partner in life.

Erklärung zur Dissertation

Hiermit versichere ich an Eides statt, dass ich die vorliegende Dissertation selbstständig und ohne die Benutzung anderer als der angegebenen Hilfsmittel und Literatur angefertigt habe. Alle Stellen, die wörtlich oder sinngemäß aus veröffentlichten und nicht veröffentlichten Werken dem Wortlaut oder dem Sinn nach entnommen wurden, sind als solche kenntlich gemacht. Ich versichere an Eides statt, dass diese Dissertation noch keiner anderen Fakultät oder Universität zur Prüfung vorgelegen hat; dass sie - abgesehen von unten angegebenen Teilpublikationen und eingebundenen Artikeln und Manuskripten - noch nicht veröffentlicht worden ist sowie, dass ich eine Veröffentlichung der Dissertation vor Abschluss der Promotion nicht ohne Genehmigung des Promotionsausschusses vornehmen werde. Die Bestimmungen dieser Ordnung sind mir bekannt. Darüber hinaus erkläre ich hiermit, dass ich die Ordnung zur Sicherung guter wissenschaftlicher Praxis und zum Umgang mit wissenschaftlichem Fehlverhalten der Universität zu Köln gelesen und sie bei der Durchführung der Dissertation zugrundeliegenden Arbeiten und der schriftlich verfassten Dissertation beachtet habe und verpflichte mich hiermit, die dort genannten Vorgaben bei allen wissenschaftlichen Tätigkeiten zu beachten und umzusetzen. Ich versichere, dass die eingereichte elektronische Fassung der eingereichten Druckfassung vollständig entspricht.

Teilpublikationen:

Gonçalves-Dias, J.*, Singh, A.*, Graf, C., & Stetter, M. G. (2023). Genetic incompatibilities and evolutionary rescue by wild relatives shaped grain amaranth domestication. *Molecular Biology and Evolution*, 40(8), msad177.

Gonçalves-Dias, J., & Stetter, M. G. (2021). PopAmaranth: a population genetic genome browser for grain amaranths and their wild relatives. *G3*, 11(7), jkab103.

José Miguel Gonçalves Dias

Köln, den 14.02.24

José Gonçalves-Dias

Bioinformatics Scientist

✉ josemgdias@outlook.com ☎ +4915258405553 📍 Frankfurt, Germany
in jmgdias 🌐 Jungal10 📄 Jose-Goncalves-Dias



📁 Professional Experience

- Bioinformatician, Universitätsklinikum Frankfurt Klinik für Kinder- und Jugendmedizin** 2023 – present
Frankfurt am Main, Germany
- Multi-omics analysis for decoding the mutational epigenetics landscape of pediatric leukemia.
 - Establishment and maintenance of the infrastructure for a cohesive data system to connect clinicians and scientists.
- Scientific Consultant, ConsultED** 2022 – present
Germany
- Personal scientific consultation and support for aspirants who want to pursue post-graduation
- PhD Computational Biology, University of Cologne** 2019 – 2022
Cologne, Germany
- Development of PopAmaranth [🔗](#), the first population genetic genome browser for plants, via hmtl and JBrowse.
Use whole-genome sequencing data from multiple populations to advance the understanding of the genetic dynamics underlying species diversification and domestication.
- Bioinformatician Internship, IKMB** 2018 – 2019
Kiel, Germany
- Use Exome-sequencing data to depict HLA expression levels in Inflammatory Bowel Diseases.
 - GWAS for multiple population patients.
- Bioinformatician Internship, DKFZ/ MRI partner site** 2017 – 2018
München, Germany
- Bioinformatics Pipeline Development for high-throughput functional cancer genomics in mice.

🎓 Education

- Ph.D. Candidate Computational Biology, University of Cologne** 2019 – present
Cologne, Germany
- "Uncovering the Genomic Mosaic of Domestication in Grain Amaranths using Computational Biology"*
- Master in Bioinformatics, IT specialization, Universidade do Minho** 2015 – 2017
Braga, Portugal
- Foundational competencies in Python, R, SQL, and MatLab for the development of algorithms for biological sequence analysis and systems biology.
Coursework and practical competence using scikit and pandas packages.
Machine Learning using decision trees.
- Erasmus LPP Exchange Program, Warsaw University of Life Sciences** 2014
Warsaw, Poland
- Bachelor Thesis, Institute of Molecular Pathology and Immunology** 2012
Porto, Portugal
- "Generation of a Vertebrate model in zebra-fish(Danio rerio) to thyroid cancer studies."
- Bachelor in Genetics and Biotechnology, UTAD** 2010 – 2015
Vila Real, Portugal
- Relevant Coursework in Biology, Genetics, and Scientific Research
- Master Thesis, Erasmus Placement Program, Chr. Hansen, A/S** 2017
Hørsholm, Denmark
- "Reconstructing the metabolic network of *Lactobacillus helveticus* on a genome-wide scale"

Scientific Publications and Presentations

Integrating Long Read Sequencing and Multiome Epigenetics for Pediatric AML Profiling

2023 *Poster*. UBC Bioinformatics Symposium

Decoding Epigenetic Landscape in Pediatric Leukemia through Multi-Omics Integration

2023 *Presentation*, XXXIV Annual Meeting of the Kind-Philipp-Foundation for Research in Pediatric Oncology, Wilsede, Germany

Genetic Incompatibilities and Evolutionary Rescue by Wild Relatives Shaped Grain Amaranth Domestication, Molecular Biology and Evolution

2023 *Publication*. José Gonçalves-Dias, Akanksha Singh, Corbinian Graf & Markus G. Stetter

PopAmaranth : a population genetic genome browser for grain amaranths and their wild relatives, G3 Genes| Genomes| Genetics

2021 José Gonçalves-Dias & Markus G. Stetter.

Genome-wide identification of Myb transcription factors in amaranth and quinoa

2019 *Poster*. Botanikertagung - International Plant Science Conference), Rostock, Germany

Skills

R

Data Manipulation, Processing, Statistical Analysis and Visualization

Expertise using tidyverse and dplyr

NGS- and 3rd gen sequencing data handling

WGS, Epigenetic (ATAC-seq, CUT&RUN/ CUT&TAG) short and long-read RNA-seq data analysis

Nextflow

high-throughput analysis and modular workflows via nf-core pipelines

HPC

HPC cloud computing using SLURM as a resource workload manager

Git/GitHub

Version control and project management

Microsoft and Google Suite

Office and videoconferencing

Awards

Travel Grant DBG Botanikertagung

2019

High School Honors Degree

2010

Local Mathematics Olympiads Winner

2006

City's Best Student

2002

Languages

Portuguese



English



German



Spanish



Extracurriculars

Population Genetics Interest Group (PPIG)

2021-2022

Bioinformatics Master Students Representative

2015-2017

Member of the Students core of Bioinformatics in UMinho (NEBIUM)

2015-2017

Member of the Permanent Board Committee of the Pedagogic Council of Environment and Life Sciences School in UTAD

2013-2014

Member and Director for the Recreational Activities section of the Academic Students Association in UTAD (aaUTAD)

2012-2014

Member of the Students Core for the Events and External Relationships Department in UTAD (adnGB)

2011-2014

Extra Activities

- Federated Handball Player
- Regular Half-Marathon Runner

Event Organization

Basics of Inflammations

2018

IV National Genetics and Biotechnology Seminar

2012

MATLAB workshop

2016

Bioinformatics Open Days

2017

VI National Genetics and Biotechnology Seminar

2014

I Bioinformatics Seminar

2012